Numerical Weather Prediction

Measuring the fit of rainfall analyses and forecasts to radar: Strategy for stage 5 evaluation

This is the stage 4 report from the Storm Scale Modelling project.
The project is half funded by Defra, project code FD2207

Forecasting Research Technical Report No. 432
Joint Centre for Mesoscale Meteorology Report No. 146

Nigel Roberts

©Crown Copyright

PROJECT DOCUMENTATION

PRODUCT DESCRIPTIONS

Measuring the fit of rainfall analyses and forecasts to radar: Strategy for stage 5 evaluation Stage 4 report from the Storm-Scale numerical modelling project

First release:
Date: 28th January 2004
Latest release:
Date: 15th April 2004

PRINCE2

Author: Nigel Roberts

Owner: Nigel Roberts

Client:

Document Number: Version 2.1

Product Descriptions History

Revision History

Date of this revision: 15/04/04
Date of Next revision:

Revision date	Previous revision date	Summary of Changes	Changes marked
	28/01/03	First issue, version 1.0	
	16/03/04	Modifications made on the basis of comments from Brian Golding 20/02/04 and 12/02/04, version 2.0	
15/04/04		Further minor corrections, version 2.1	

Approvals

This document requires the following approvals. Signed approval forms are filed in the Quality section of the PCB.

Name	Signature	Title	Date of Issue	Version
Brian Golding	*Brian Golding*	Hd(FR)	30/6/4	2.1

Contents

1 Introduction 5
1.1 General 5
1.2 What is it we want to know 5
2 Current data assimilation 6
2.1 The analysis time is really an analysis period 6
2.2 Data assimilation methods 7
2.2.1 Three-dimensional variational data assimilation (3DVAR) 7
2.2.2 Moisture Observation Pre-Processing System (MOPS) – Latent heat nudging 8
2.2.3 MOPS cloud analysis 9
2.2.4 Addition of low-resolution 3DVAR increments 10
3 The verification approach 11
3.1 Decisions 11
3.1.1 Verify precipitation accumulations 11
3.1.2 Verify against radar data 11
3.1.3 Evaluate the accuracy of the locations at which thresholds are exceeded 12
3.1.4 Evaluate at a range of scales coarser than the grid scale 13
3.2 How fractional exceedance is computed 14
3.3 Verification measures for fractions/probabilities 16
3.3.1 The Brier skill score 16
3.3.2 Skill score for area fractions 18
3.3.3 The Odds ratio 19
3.3.4 The reliability diagram 21
3.3.5 The ROC curve 23
3.4 How to deal with radar error? 24
4 Idealised verification examples 25
4.1 Example 1. 26
4.2 Example 2 30
4.3 Example 3 32
4.4 Example 4 34
4.5 Example 5 37
4.5.1 Conclusions from the idealised examples 40
5 Intended assessment 41
6 Appendix A 43
7 Appendix B 44
8 References 45

1 Introduction

1.1 General

This is the 4th interim report to be delivered from the Storm Scale Numerical Modelling project. The project was set-up to investigate the ability of a storm scale configuration of the Met Office Numerical Weather Prediction (NWP) model to predict flood-producing rainfall events. In this context a storm-scale model has a horizontal grid spacing of about 1 km and receives information at the boundaries from a model with 4-km grid spacing.

The report will describe a strategy for examining the fit of rainfall analyses and forecasts to radar data. In the next stage of the project (stage 5) the methodology described here will be used examine the performance of 12, 4 and 1km gridlength model simulations of convective events.

The initial intention was to examine the fit of model analyses alone to radar in this stage and then move on to look at forecasts separately in stage 5. It is now thought to be more sensible to evaluate the fit of analyses and forecasts together, because of the way current data assimilation systems introduce observations into models.

1.2 What is it we want to know

All numerical model forecasts should start by representing the state of atmosphere as accurately as possible. The start time of a forecast is called the analysis. The way to make a numerical model more closely represent the true atmospheric state at the analysis time is to introduce observations into the model through a process called data assimilation. Data assimilation involves taking a previous forecast for a few hours ahead and then blending in the information from new observations in a way that makes the model forecast agree as closely as possible with the new observations, but without making the model susceptible to the generation of spurious noise when a new forecast is run. See Figure 1. This process must take account of inherent errors in observations as well as errors in the previous model forecast.

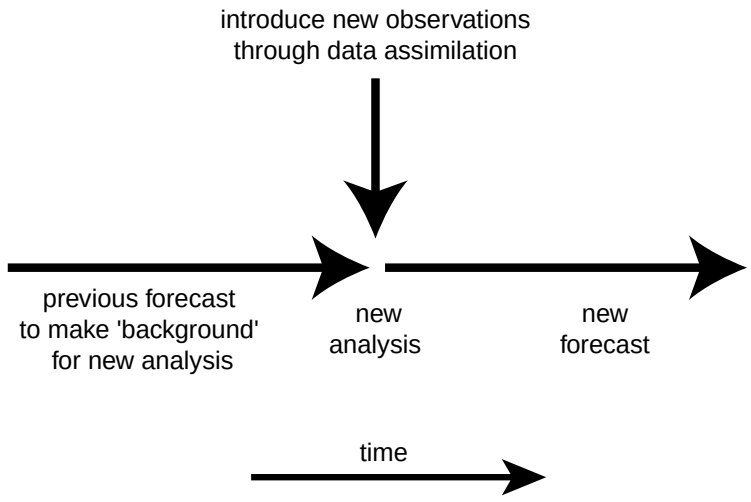

Figure 1. Schematic diagram of the process of introducing data into a numerical model at the analysis time.

The previous forecast, at the time into which new observations are introduced, is known as the 'background' and should already be in reasonable agreement (at scales that can be properly resolved) with the new observations if the forecast period is short. In an operational system, the introduction of observations followed by a short forecast followed by the assimilation of more up-to-date observations is a continuous process. Each time new observations are assimilated another 'assimilation cycle' has been completed. Forecasts are also continued beyond the time required to be used as a background so that weather forecasts can be produced for customers.

It is fundamentally important to assess regularly how closely analyses and forecasts fit observations so that the performance of data assimilation systems and forecast models can be monitored. For the operational Unified Model system this is done routinely to check performance.

This project is concerned with assessing the capability of a new storm-scale version of the Unified Model to predict high-impact precipitation events. For such a system, it is also vital to examine how well the model analyses and forecasts fit observations. In particular, we wish to: -

1. Quantify how well data assimilation improves model precipitation at the analysis time – and compare different data assimilation techniques
2. Quantify whether better analyses result in better precipitation forecasts - and how quickly the skill drops off with forecast time

Two main problems that arise when attempting to do this

1. All of the currently available data assimilation methods update the model over a period of time rather than at a single moment in time. This means that in reality it is difficult to define such a thing as an analysis time.
2. There are particular difficulties associated with the evaluation of precipitation from a high-resolution model.

Data assimilation methods will be discussed in section 3.2. The approach to precipitation verification will be discussed in section 4.

2 Current data assimilation

There are four methods currently available at the Met Office for introducing observations into a high-resolution numerical model.

1. Three dimensional variational data assimilation (3DVAR)
2. Moisture Observation Pre-Processing System (MOPS), Latent heat nudging
3. MOPS cloud analysis
4. Interpolation of 3DVAR increments from a coarser resolution model

They will be briefly described in section 4.2 after some words about how the concept of an analysis time should really be thought of as an analysis period.

2.1 The analysis time is really an analysis period

Figure 1 shows a simple schematic of an assimilation cycle with the background forecast being adjusted to the observations at the analysis time. Unfortunately this is not exactly how it works in reality. Figure 2 gives a better impression. The observations are not introduced into the model at a single analysis time; they are incorporated at intervals over

a period of time during a background forecast. This is the case for the current operational system and may be for a high-resolution modelling system. As a result it is difficult to define a single analysis time. The most appropriate time to take as the analysis might appear to be when all the observations have been included, but this may be a considerable time after the background forecast has started. Also, the simple picture hides the fact that there may be several different techniques for assimilating different types of observations going on at the same time and the period over which these methods operate is different. The end of the assimilation period will depend on the assimilation method(s) used. A nominal 'analysis' time is taken to be at a time called T+0 within the middle of the assimilation period. It defines the start time of the next background forecast even though more observations are still to be added. For the operational mesoscale model (gridlength 12km) the assimilation cycle is 3 hours and the 'analysis' times are taken to be 00,03,06,09,12,15,18 and 21UTC each day. Observations are introduced through the data assimilation process around each analysis time.

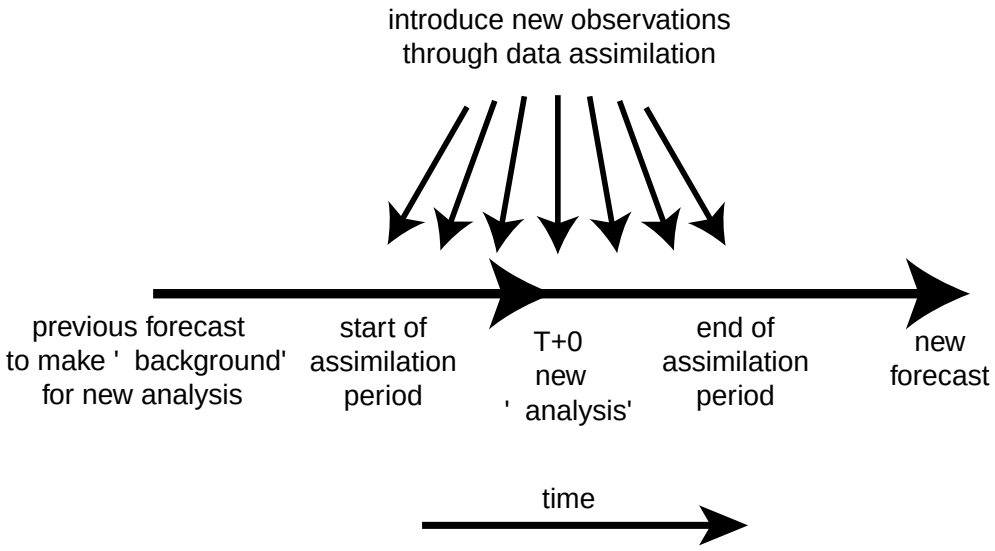

Figure 2. A schematic diagram of the process of introducing data into a numerical model over an analysis period. More realistic than Figure 1.

For the purposes of evaluating the fit of high-resolution model precipitation fields to observations, it makes much more sense to examine how the fit varies over an assimilation period and on into the subsequent forecast, than to call T+0 the single analysis time and all that follows the forecast. It is expected that the fit will improve through the analysis period if the data assimilation is working satisfactorily.

2.2 Data assimilation methods

2.2.1 Three-dimensional variational data assimilation (3DVAR)

3DVAR is the technique that is used to adjust the model to be closer to observations of pressure, temperature, humidity and wind. It is used operationally in the 12km-gridlength mesoscale model (and the global forecast model, see Lorenc et al 2000). It is now available to be tested in a model with a gridlength of 4km, but is currently too costly to run with a gridlength of 1km.

The aim of 3DVAR is to minimise the misfit of a model state at a particular time to new observations and to a background state. The background state comes from a previous forecast (discussed earlier). The model state is a reasonable guess, which to start with is usually the same as the background. The variables that define the model state are varied until a compromise is reached in which the combined misfit to both the observations and the background state is minimised. This process is crucially dependent on the expected errors in the background state and the observations and also on the influence that each model gridpoint is expected to have on every other. Once the new model state (the analysis) has been found the differences between this new analysis and the background state are found and stored as increments.

These increments are then added to the background model forecast bit by bit through a period of time as shown in Figure 2. The reason for adding the increments in this gradual way is to enable the model variables to mutually adjust over a period so that noise is not generated by the shock. The time period used to add the increments from 3DVAR in the operational 12km gridlength forecast system is 2 hours (see Figure 3).

The types of observations used are:

1. Surface reports of surface pressure, temperature, humidity and wind, from conventional surface stations, buoys and ships and wind profilers.
2. Radiosondes – that provide vertical profiles of temperature wind and relative humidity.
3. Aircraft measurements of temperature and wind.
4. Satellite measurements of winds by tracking cloud features in imagery
5. Satellite measurements of vertical profiles of temperature and humidity

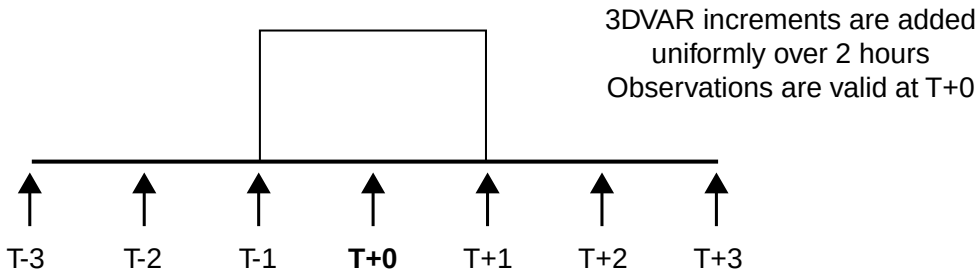

Figure 3. A schematic of the weighting given to observation increments added through 3DVAR. Courtesy of Zhihong Li.

2.2.2 Moisture Observation Pre-Processing System (MOPS) – Latent heat nudging

Latent Heat Nudging (LHN) (Jones and Macpherson1997) is a technique that has been developed to adjust the precipitation in a numerical forecast model to be closer to the observed precipitation. The method is used operationally in the 12km-gridlength mesoscale model. The rain rates are not modified directly. Instead, temperature increments are added to the model, to scale the model latent heating rate profile (heat that is generated from condensing water vapour into cloud and precipitation) by the ratio of observed rain rate to forecast rain rate. If the model is not producing rain where rain is

observed, then a nearby point is used to supply a heating rate profile for scaling. The idea is, that once the temperature profiles have been adjusted, the model winds and vertical motion will adjust accordingly to produce a more realistic rainfall pattern (rain or snow results from saturated ascent).

The observed rain rates come from Nimrod (Golding 1998) processed radar data and are smoothed to the model grid if necessary. The weighting given to MOPS latent heat increments is shown in Figure 4.

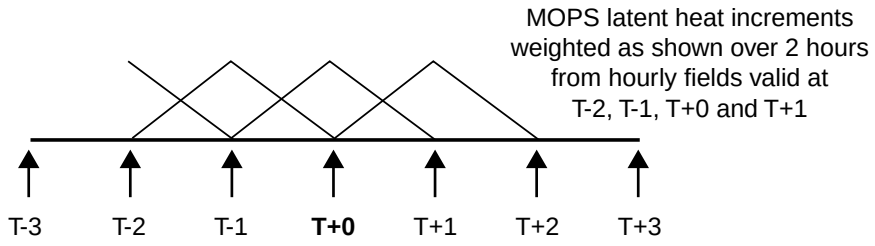

Figure 4. A schematic of the weighting given to observation increments added through latent heat nudging. Courtesy of Zhihong Li.

2.2.3 MOPS cloud analysis

This is a technique for adjusting the cloud in a model to give a better fit to the observed cloud structure (Macpherson 1996). It is used operationally in the 12km-gridlength mesoscale model. Nimrod radar analyses, surface rainfall measurements, Meteosat images and surface cloud cover reports are combined with the model background, to produce a 3-dimensional picture of the 'actual' cloud amount in terms of cloud fractions within the model domain. The cloud fractions are then converted into profiles of relative humidity by use of the Smith scheme (Smith 1990) to relate cloud fraction to grid-box mean relative humidity. The humidity profiles are then nudged as increments into the model over a period of time. If both the MOPS cloud fraction and a model grid square have no cloud, then no nudging is done. Humidities are converted into cloud fraction within the model using the Smith scheme. The weighting given to MOPS latent heat increments is shown in Figure 5.

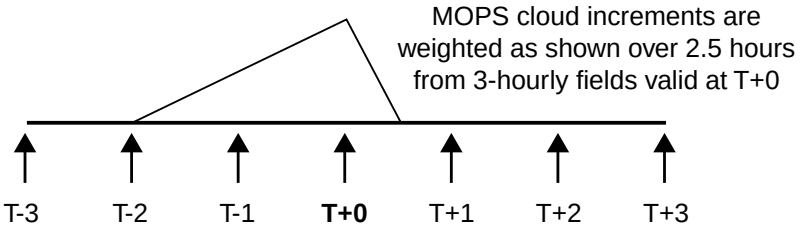

Figure 5. A schematic of the weighting given to observation increments added through latent heat nudging. Courtesy of Zhihong Li.

2.2.4 Addition of low-resolution 3DVAR increments

Unlike the other methods, this approach is not used operationally. It is used specifically for high-resolution modelling because 3DVAR becomes increasingly computationally expensive as resolution is improved.

3DVAR is performed at 12km or 4km gridlength to produce a set of increments. These increments are then added bit by bit to the background state of the higher resolution model we are interested in. For example, 3DVAR might typically be performed on a 4-km gridlength 3-hour forecast (the 4-km background) to get a set of increments to add to that background in order to create a new 4-km analysis. Instead, these increments are added bit by bit to a 1-km gridlength background (an equivalent 1-km gridlength 3-hour forecast) in order to create a new 1-km analysis. The benefit of doing this is that a high-resolution forecast can be run without either having to start with fields interpolated from a coarser resolution model or having to run with 3DVAR at 1-km gridlength. It provides a way of preserving the high resolution detail from the previous forecast while adjusting the larger scales to match the observations. There are however two drawbacks. (1) The information being added to the high-resolution model is not at the same resolution as the high resolution model, so important detail may be lost. (2) The coarser-resolution background forecast used by 3DVAR is different from the higher-resolution background forecast the increments are added to. This may mean that, in time, the two different resolution models could drift apart when they are supposed to be giving similar broad-scale forecasts. The mismatch may also generate spurious noise in the high resolution forecast.

3 The verification approach

3.1 Decisions

Before going ahead with the verification of precipitation forecasts, several decisions need to be made about the way the verification should be done. Four decisions are presented in this section.

1. Verify precipitation accumulations rather than precipitation rates.
2. Verify rainfall against radar rather than use rain gauges.
3. Evaluate the accuracy of the locations at which thresholds are exceeded rather than the amount of accumulation at each location, so as to avoid the difficulties with a mixed distribution of zeroes and non-zero amounts
4. Evaluate at a range of scales coarser than the grid scale and select the appropriate scale to use in order to optimise the benefit.

3.1.1 Verify precipitation accumulations

Both rainfall rates and accumulations can be generated by numerical forecast models and estimated from radar. Precipitation rates give an instantaneous picture of a rainfall event, whereas accumulations show a temporal average. The decision has been made to verify only precipitation accumulations at this stage. The reasons for this are:

1. In terms of forecasting extreme rainfall events, it is usually the precipitation accumulations that matter rather than the distribution of the rain at any specific moment in time.
2. If the precipitation output is being smoothed spatially by converting into probabilities then it also makes sense to smooth over time.
3. Rainfall rates are very noisy. Peak values may be very small-scale, unpredictable and transient, especially in convective situations. Verification of instantaneous pictures may give misleading results unless a large number of times are examined – and this is in effect, the same as using accumulations.
4. Errors will be present in the radar data used for the verification. Some of these errors will be sporadic and their amplitude may be reduced by looking at accumulations over a sufficiently long time period.
5. Fast moving features like rapidly propagating fronts may have large displacement errors if viewed at a single time, but this might be a relatively small timing error and not significant in terms of the amount of rain that is produced.

3.1.2 Verify against radar data

For the verification of precipitation there are three options

1. Compare model precipitation with rain gauge data
2. Compare model precipitation with radar data
3. Do both of the above.

The decision has been made to compare with radar data only at this stage.

The radar data is a composite of several radars across the UK. Gauge data is quantitatively more accurate at a single point, but there are serious representivity problems. Even if the

rainfall is accurately measured at a particular point, it may be a completely different just a few hundred metres away. The only way to get a satisfactory view of the spatial distribution of rainfall is to have an extremely dense network of rain gauges over a large area – and such a network does not exist. The other problem with using gauge data is that NWP models produce rainfall that is an average over a gridsquare rather than a point value – it is not a like for like comparison. To convert gauge data into gridsquare averages would be possible if there were a sufficiently large number of gauges in each gridsquare, but since we are aiming to verify high-resolution model output defined on grid squares, a phenomenally dense network of gauges would be needed.

Radar data, on the other hand, can provide a view of the spatial distribution of precipitation that gauges can not. Rainfall is presented as averages over 5x5 km squares that encompass most of the UK (and over 1x1km squares for smaller areas) so can be compared directly with model output in a like for like way. It is not as accurate as gauge data. The difference between the radar and gauges is estimated to be up to a factor of 2 (Harison et al) – however much of this difference is due to representivity errors (in space for gauges, and in time for the radar). Compared to some (unknown) truth, radar error is thought to be less than a factor of 2, but the difficulty is that it is not properly known.

The deciding factor in choosing to use radar data is the advantage of being able to see the spatial distribution of precipitation. This will allow us, in principle, to evaluate the ability of a model to produce the correct rainfall coverage and patterns and determine spatial error. Clearly, the errors in the radar data need to be accounted for in some way (both in magnitude and displacement), even so, it is fair to say that, if a high-resolution forecast system consistently produces precipitation that looks very similar to what the radar 'sees', it can be regarded as a very skilful forecast system.

The radar data we intend to use is a composite from the UK rainfall radar network. It has been corrected for gross errors such as spurious signals from the beam hitting the ground or other objects (anaprop) and the increased return signal from melting snow (bright band).

A product in which the radar data and gauge measurements are merged is now being development. This will be the preferred choice in future.

3.1.3 Evaluate the accuracy of the locations at which thresholds are exceeded

This helps to deal with the problem of having mixed distributions. If accumulations are directly compared with accumulations there will be a problem at grid squares where either the forecast or the observation (or both) are zero. The distribution is approximately log normal, but it is impossible to compute differences in log accumulation because the zeroes become undefined. On the other hand, computing differences in accumulation might be inappropriate as it will weight errors in large amounts – forecasts of high accumulations will give larger errors even if the forecasts are relatively good.

The use of a threshold gets rid this problem by giving each point over the threshold the same value. Pixels (grid squares) exceeding the threshold are given a value of 1 and all others a value of 0 (or a missing data indicator if the radar does not 'see' that pixel). Since all accumulations over the threshold will be set to a value of 1 regardless of the magnitude of the accumulation, several thresholds will be required to evaluate errors in magnitude as well as spatial errors. For example, if 4 mm was the threshold and most of the forecast pixels exceeding the threshold had values over 10 mm, but the radar pixels exceeding the threshold rarely exceeded 6 mm, the way to discover this over forecasting would be to use an additional threshold of 8 mm.

3.1.4 Evaluate at a range of scales coarser than the grid scale

The third interim report from this project explains why we should not believe precipitation features at or near the grid-scale of a high-resolution model. A 1 km-gridlength model cannot achieve accuracy to 1 km or even a few kilometres because features are not resolved properly numerically and there is very low predictability on those scales. Showers can not usually be predicted in exactly the correct locations even if a forecast or analysis is good. The displacement errors may be comparable to the size of the showers themselves and this can lead to very poor verification scores – and an apparently very poor forecast. So the implication is that it is possible to have what is thought to be a 'good' forecast, yet have poor verification scores – but what is meant by 'good' and how can it be measured? What is usually meant is that, although the detail is not exactly correct, the region of shower activity, or the frontal zone, or storm complex is well positioned and has the right type of structure and intensity. We judge forecasts by eye over a variety of scales by saying that they are good at getting a large feature like a front in the right place, but are not quite so accurate for organised thunderstorms and poorer still for individual showers. But how can we determine this objectively? Traditional methods for evaluating model performance have generated statistics or scores on a gridpoint by gridpoint or observation point by observation point basis alone. That is not a viable approach here because we do not expect the model to be skilful at the grid scale and it says nothing about skill over any other scales. Somehow we must be able to assess the fit of model precipitation output to radar data over different spatial scales so that the performance of models and data assimilation methods can be examined over different scales and appropriate scales can be found for precipitation diagnostics that are to be presented to customers.

We would like to at least attempt to answer the following questions

1. How accurately can we forecast precipitation over areas the size of counties, river catchments or urban areas using a particular model?
2. How does the predictable scale change with forecast time? This may be in terms of defining the smallest catchment area for which forecasts are useful.
3. Does the rainfall analysis agree with the radar picture at the scale of the data assimilation? Data assimilation methods are designed to add observational information to a model over particular spatial scales.
4. What are sensible products to generate for customers?
5. Does a change to either model resolution or formulation make a difference to the predictable scale?

This list of questions further justifies the need to verify over scales larger than the grid scale. This will be done using two approaches. Both of these involve computing the fraction of grid squares in an area that exceed a particular accumulation threshold. The simple methodology for doing this is explained in section 4.2.

(a) Verification approach 1. For areas of different size, compare the fraction of the area that exceeds the threshold in both forecast and observation.
(b) Verification approach 2. For areas of different size, use the fractional area that exceeds the threshold to define a probability of exceedance at the grid scale, and then compare with the true probability at the grid scale (i.e. either 0 or 1)

Fractional coverage can be thought of as a probability. The generation of probabilities for output diagnostics and products was described in the third interim report. It is vital to link model evaluation with model products and this approach provides a way of doing just that.

3.2 How fractional exceedance is computed

A fractional area can be generated by finding the fraction of neighbouring points surrounding each gridpoint that have a rainfall accumulation above a particular specified threshold. A point can be defined as neighbouring another gridpoint if it falls within a square region surrounding that gridpoint. Different sized squares can be used to produce different pictures – a small square means that only information close to each gridpoint is used, whereas the use of a larger square will take information over a larger area around each gridpoint and produce a smoother field.

Figure 6(a) gives an example of how fractions can be calculated over different sized squares for a particular rainfall accumulation threshold. In this example, different fractions are computed for the point marked with the Red Cross when the size of the square is changed. The blue crosses represent points where the rainfall accumulation threshold is exceeded and the black crosses are points where that threshold is not reached. The use of the three different sized squares gives the following fractions.

Square 1	3x3 points	Fraction = 2/9	= 0.222
Square 2	7x7 points	Fraction = 21/49	= 0.429
Square 3	13x13 points	Fraction = 44/169	= 0.260

The fractions need not necessarily be generated within squares. Circles could be used instead. Points could be weighted according to their distance from the centre of a square or circle. For the purposes of this verification, it is simpler to use squares and a uniform weighting at this stage.

Model precipitation accumulations can therefore be compared with radar accumulations over different spatial scales by the use of these fractions. As indicated in the previous section, there are two ways this can be done:

Verification approach 1 – Comparison of fractions with fractions

Now imagine that Figure 6(a) is from a model forecast and Figure 6(b) is from radar with no error. They are both similar in overall pattern, but with some differences in detail. The probabilities for the point with the Red Cross in Figure 6(b) over the three squares are: -

Square 1	3x3 points	Probability = 7/9	= 0.778
Square 2	7x7 points	Probability = 25/49	= 0.510
Square 3	13x13 points	Probability = 43/169	= 0.254

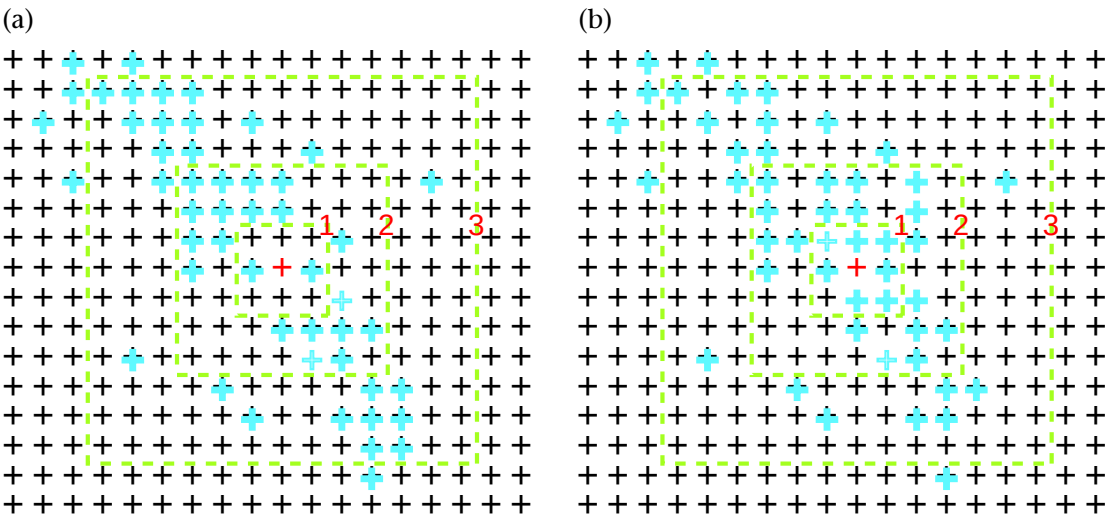

Figure 6. A schematic of how probabilities are generated using different sized squares. Crosses represent gridpoints. Green dashed lines show squares centred around the Red Cross. The blue/black crosses represent gridpoints with a precipitation accumulation above/below a threshold. (a) is for a forecast and (b) for radar. See text for more information.

The interpretation of these fractions is that, if you were to pick a point at random in the small square you would have a 22% chance of getting a rainfall accumulation over the threshold in the forecast and a 78% chance in reality. There is clearly a large discrepancy between the two. However, take the large square and there is a 25% chance in the forecast and a 26% chance in reality - now the chances are almost the same. The near agreement over the large square indicates that the forecast is good on that scale, but over the smaller scale the performance is much worse. This agrees with the assessment by eye that the general patterns are similar but there are differences in detail. Of course, if the same was done for squares centred on a different point the result might be quite different. To produce a useful signal, the probabilities over different sized squares should be compared over all points in the domain.

Verification approach 2 – Comparison of model fractions with grid points

Forecast fractions can be compared against a binary radar field in which all points exceeding the accumulation threshold have a value of 1 and all others are 0. Clearly, the larger the squares over which the fractions are computed, the less likely it is that those fractions will be close to either 1 or 0. Fractions generated over very small squares are more likely to discriminate and have values close to 1 or 0 but may be at the wrong locations. The best spatial scale for a particular is forecast will be the size of the square that makes the best compromise between these two competing effects.
The choice of method depends on the type of verification score being used and on the information we wish to gain. Verification scores will be discussed shortly.

If fractions are to be generated from radar accumulations, then the process needs to take account of missing data. Data will be missing either because points are too far away from the radars or because there was a problem with the data. Fractions are only computed at a gridpoint for a particular sized square surrounding that point, if a sufficient proportion of points in the square have data – otherwise that gridpoint is not used in the verification. The other reason for omitting some points from the verification is if they are within a few gridlengths of the edge of the domain. Boundary effects in a numerical model mean that points close to the boundary should be discarded anyway when presenting a forecast. In generating the probabilities for points close to the boundary the squares can fall over the edge of the boundary and significantly reduce the sample size.

The fractions can be interpreted as probabilities and presented to customers in that way. The use of probabilities for products is discussed in the stage 3 report.

3.3 Verification measures for fractions/probabilities

The verification measures that will be used and described here are:

1.	The Brier skill score	approach 2, for fractions vs binary
2.	Fractions skill score	approach 1, for fractions vs fractions
3.	The odds ratio	primarily approach 1, for fractions vs fractions
4.	The reliability diagram	approach 2, for fractions vs binary
5.	The ROC curve and area	approach 1, for fractions vs fractions

Idealised results will be shown later for the Brier skill score, fraction skill score and odds ratio. Other measures will also be mentioned. Each score has different characteristics and will measure different aspects of the model performance. No single score is sufficient to provide all the information we need.

3.3.1 The Brier skill score

Firstly the Brier score needs to be defined. It is the mean of the sum of the squares of the differences between the forecast probabilities and observed points. The observed points have the value 1.0 where an event occurs and 0.0 where an event does not occur. It is in effect, the mean square error for probabilities and is used to examine the magnitude of the probability errors.

$$BS = \frac{1}{N} \sum_{j=1}^{N} (p_j - o_j)^2$$

$0 \leqslant p_j \leqslant 1$ probabilistic forecast

$o_j = 1$ if event occurs

$o_j = 0$ if event doesn't occur

$N =$ total number of points

For a perfect forecast BS = 0.0. The only way this can happen is if the only probabilities to occur are $p_j = 0.0$ or 1.0 and all the occurrences are correctly predicted. The worst possible score is 1.0.

The problem with the Brier score is that it is sensitive to the climatological frequency of an event being forecast (and therefore the frequency of the event itself). If a forecast system

consistently fails to forecast rare events it will still get a good score. It could favour a cautious forecast system over a more ambitious and useful forecast system. Since we are interested in how well forecasts warn of extreme events, this is not a desirable property.

The sensitivity to the frequency of the event can be largely removed by constructing a Brier skill score. It is used to find the relative skill of the probabilistic forecast over that of a reference forecast or climatology.

$$BSS = \frac{BS - BS_{reference}}{BS_{perfect} - BS_{reference}} = \frac{BS - BS_{reference}}{0 - BS_{reference}}$$

$$BSS = 1 - \frac{BS}{BS_{reference}}$$

The Brier skill score has a range of minus infinity to 1.0, with 1.0 for a perfect forecast and a value of less than 0.0 for a forecast with less skill than the reference forecast.

The reference Brier score $BS_{reference}$ could be a climatological value – this would need to be determined from a large number of events, or it could be another forecast – e.g. to compare one resolution against another. Another option is to take as the reference Brier score the Brier score that would be obtained if the same number of points that exceed the accumulation threshold in the radar picture were scattered randomly over the domain and compared with the radar picture for that accumulation. We are asking if the model can do better than a random field with the same frequency as the observed radar field. The Brier score for a random field is given by:

$2f(1-f)$ f = the observed frequency

Therefore the Brier skill score will be

$$BSS = 1 - \frac{BS}{2f(1-f)}$$

In principle, this Brier skill score used with probabilities of model rainfall accumulations generated over different sized squares can be used to examine how well the model performs over different spatial scales but there are problems.

One problem is that when forecast probabilities are generated over different sized squares, the score is dependant on whether the observed frequency is larger than the forecast frequency or the other way round. This is not desirable because it means that we can not rely on this score to consistently tell us anything about spatial error in a forecast – as the same spatial error could give two different scores. The other problem is the use of the Brier score for a random field as a reference forecast. To do this makes an assumption that we know in advance that the random forecast will have the same frequency of precipitation as that observed. We are assuming prior knowledge we do not have and giving our reference forecast a very high level of inherent skill – a level of skill that is difficult to beat.

A way to get round both of these problems is to make the forecast and observed frequencies the same before doing the verification. This can be done by ranking the gridpoints in the model and radar fields by increasing precipitation accumulation and then compare the points in say the top 10% of one with the points in the top 10% of the other. A set of suitable percentage values could be chosen as thresholds, e.g. 50, 75, 90, 95, 99%. By definition, the random-field Brier score does not now assume any prior knowledge (as we have fixed the frequency we want) and it is not an unrealistically skilful forecast since it can not do better than the actual forecast in terms of the frequency of the event. The problem with this approach is that the accumulation threshold being compared will not be the same (the model 90% value might be 2mm, whereas the radar 90% value might be 5mm). However, the positive aspect is that this difference does provide an indication of the bias.

3.3.2 Skill score for area fractions

As well as using the Brier score to compare forecast probabilities with binary observed probabilities, we can use the same method to compare forecast area fractions (or probabilities) with observed area fractions (or probabilities) and get a 'Brier score' for fractions $BS_{fractions}$.

$$BS_{fractions} = \frac{1}{N}\sum_{j=1}^{N}(p_j - o_j)^2 \qquad 0 \leqslant p_j \leqslant 1 \quad \text{forecast area fraction}$$

$$0 \leqslant o_j \leqslant 1 \quad \text{observed area fraction}$$

As for the standard Brier score, a perfect forecast gives a $BS_{fractions} = 0.0$. The worst possible score is 1.0.

There are, however, some drawbacks to doing this:

1. It has the same problem as the standard Brier score in being sensitive to the climatological frequency of an event being forecast. As for the standard Brier score, it is better to convert to a skill score instead.
2. If we wish to use this score to examine the skill of forecasts over different sized squares, then we run into a difficulty. Imagine a situation in which a forecast produced a single grid square of rain in one location when the radar had a single grid square of rain 10 grid-lengths away. If this score is computed for fractions over squares the size of the grid squares (at the gridscale) the score will be 2/N where N is the number of grid squares. If the score is again computed for fractions generated over 3x3 grid-length squares, it will come out as 2/9N, which is a better score, even though there is no more useful skill at that scale than there was at the grid scale. We should only really see an improvement in this example if the square size used is 10 grid lengths or larger. The score will show an apparent increase in skill with spatial scale whether there really is or not.
3. There is a problem with comparing fractions. We could ask the question: Is the difference between 0.0 and 0.2 the same as the difference between 0.4 and 0.6? With this score it is. However, in reality we know that there is a very large difference between zero chance of rain and a 1 in 5 chance and very little diffference between a bit less than a 1 in 2 chance and a bit more than a 1 in 2 chance.

The first two problems can be removed by constructing a skill score in which the Brier score for area fractions is compared to the worst possible score that could be achieved given the same forecast and observed frequency.

$$\text{Fractions skill score} = 1 - \frac{BS_{\text{fractions}}}{\text{Worst score}}$$

In effect, it is computing the score that would be obtained if there was no co-location of any of the forecast and observed grid squares exceeding a particular threshold. It also means no co-location of any forecast and observed fractions computed using a particular sized square.

$$\text{Worst score} = \frac{1}{N}\left[\sum_{j=1}^{N}(p_j)^2 + \sum_{j=1}^{N}(o_j)^2\right] \qquad \begin{array}{ll} 0 \leqslant p_j \leqslant 1 & \text{forecast area fraction} \\ 0 \leqslant o_j \leqslant 1 & \text{observed area fraction} \end{array}$$

The characteristics of this skill score are:

1. The range is 0.0 to 1.0
2. The score for a perfect forecast in which the forecast and observed rain areas (exceeding a threshold) exactly match is 1.0
3. The score for a forecast which has no rain over the threshold in the correct place is 0.0
4. If no rain over the threshold is forecast but some occurs the score will be 0.0
5. If rain is forecast but none occurs the score will be 0.0
6. If the frequency (the fraction over the whole domain) of the forecast and observed rain is the same the score will tend to 1.0 for fractions computed over very large squares.
7. If the frequencies of the forecast and observed rain are different then the score for fractions computed over larger and larger squares will asymptote to a value that can be determined from the frequencies (see appendix B).
8. The score is more sensitive to events with small frequencies because it is sensitive to the ratio of the frequencies. The score it will asymptote towards for increasing fractional will be worse if the observed and forecast frequencies are 0.1 and 0.2 (ratio of 2) than if the frequencies are 0.2 and 0.3 (ratio of 3/2). This is not a bad characteristic as it deals with the 3rd of the problems described earlier.
9. If a frequency threshold is used instead of an accumulation threshold, so that say the top 5% of the observed fractions are compared with the top 5% of forecast fractions, then the score will be purely determined by spatial errors in the forecast and will asymptote to 1.0 for large fractional areas.

3.3.3 The Odds ratio

The odds ratio measures the ratio of the odds of making a hit to the odds of making a false alarm (see Stephenson 2000 for more information). It examines how much we should believe a forecast. It is a categorical score that is calculated from the elements in a contingency table.

Contingency table

A particular event (e.g. rainfall accumulation exceeding a threshold) can either be forecast or not forecast and either observed or not observed at each gridpoint. A 2x2 contingency table is used to split the total number of gridpoints into the four possible combined outcomes: -

	Observed Yes	Observed No
Forecast Yes	A	B
Forecast No	C	D

A = number of hits (correctly forecast events)
B = number of false alarms (incorrectly forecast non-events)
C = number of misses (incorrectly forecast events)
D = number of correct rejections (correctly forecast non-events)

The odds ratio is calculated from a contingency table as AD/BC (see below).

$$\text{Odds} = \frac{p(x)}{p(\bar{x})} = \frac{p}{1-p} \qquad p = \text{Probability of something}$$

$$\text{Probability of a hit} \quad H = \frac{A}{A+C}$$

$$\text{Probability of a false alarm} \quad F = \frac{B}{B+D}$$

$$\text{Odds (hit)} = \frac{A}{A+C}\left(1-\frac{A}{A+C}\right)^{-1} = \frac{A}{C}$$

$$\text{Odds (false alarm)} = \frac{B}{D}$$

$$\text{Odds ratio} = \frac{\text{Odds (hit)}}{\text{Odds (false alarm)}}, \theta = \frac{\frac{H}{(1-H)}}{\frac{F}{(1-F)}} = \frac{AD}{BC}$$

The odds ratio has a range of 0 to infinity; 1 indicates no skill. A perfect score is infinity. It is common to plot logarithm of the odds ratio, then the log odds ratio has a range from minus infinity to infinity; 0 indicates no skill. A perfect score is infinity.

The main advantage of the odds ratio over many other scores is that it takes into account the frequency of the event. However, it does have a tendency to give better scores for rarer events. A difficulty is that there may be occasions when one or more of the values A, B, C or D in the contingency table is equal to zero, in which case the odds ratio is undefined.

The biggest obstacle to using the odds ratio is that we wish to verify probabilities and the odds ratio is calculated from categorical 'yes' or 'no' outcomes. The probabilities have to be converted into the categorical outcomes by using a probability threshold. Then a contingency table can be set up such that: -

Hit (A) = forecast probability > threshold, radar probability > threshold
False alarm (B) = forecast probability > threshold, radar probability < threshold
Miss (C) = forecast probability < threshold, radar probability > threshold
Correct negative (D) = forecast probability < threshold, radar probability < threshold

However, the most appropriate probability threshold to use is unknown because it varies from case to case and the choice of probability threshold can make a difference to the score. One approach is to use a large number of probability thresholds and build a separate contingency table for each one, then add all the contingency tables together before calculating the odds ratio. This appears to give sensible and useful values when applied to idealised cases – discussed in section 4.

The odds ratio could be presented in the following ways:
1. On a plot against increasing square sizes, for a given threshold and accumulation period – to examine how the forecast skill varies with spatial scale.
2. On a plot against forecast time, for a given threshold and spatial scale – to examine how the skill varies with forecast time.
3. On a plot against accumulation threshold, for a given spatial scale and accumulation period.

This is not an exhaustive list and suitable presentation techniques will need to be developed.

It is important to make the distinction between the odds ratio and the ratio of the probability of making a hit to the probability of making a false alarm. The latter is saddled with a dependence on the frequency of an event.

3.3.4 The reliability diagram

The reliability diagram is used to compare the forecast probabilities of an event occurring with the fraction of occasions when the event did occur. It is hoped that they should agree over a large enough sample if the forecast system is good. The use of the reliability diagram is shown in verification by Hand 2002.

The way this is done is to group the forecast probabilities into eleven probability bins to represent probabilities of 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100%.

1. 0.00 to 0.05 probability = 0 %
2. 0.05 to 0.15 probability = 10 %
3. 0.15 to 0.25 probability = 20 %
4. 0.25 to 0.35 probability = 30 %
5. 0.35 to 0.45 probability = 40 %
6. 0.45 to 0.55 probability = 50 %
7. 0.55 to 0.65 probability = 60 %
8. 0.65 to 0.75 probability = 70 %
9. 0.75 to 0.85 probability = 80 %
10. 0.85 to 0.95 probability = 90 %
11. 0.95 to 1.00 probability = 100 %

The number of occasions on which an event occurred is counted for each probability bin. The event will be a particular accumulation threshold being exceeded.

The reliability diagram is then plotted as a graph of forecast probability on the x-axis against the fraction of occasions on which the event occurred on the y-axis.

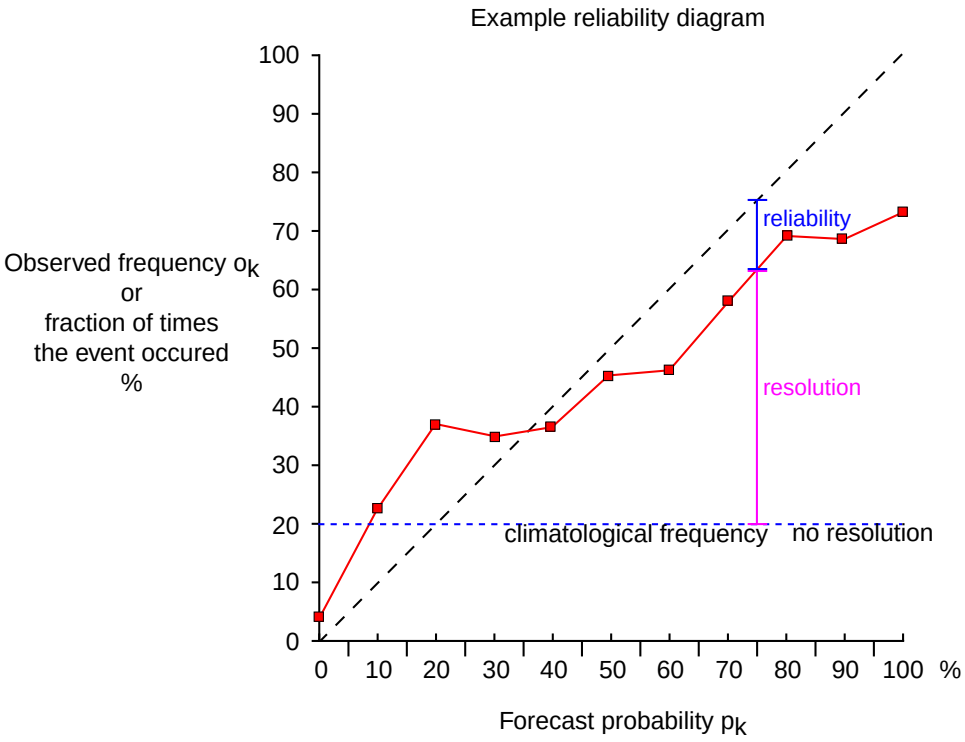

Figure 7. Example of a reliability diagram. See text for more information.

For a forecast with perfect reliability the red line will lie along the diagonal dashed black line. So for example, if a probability of between 25 and 35% for an event occurring is forecast 1000 times, then there should be around 300 occurrences of the event (frequency 30%). If the line lies above the diagonal there is under-forecasting (the probabilities are too low), if it is above there is over forecasting (the probabilities are too high). The flatter the line the lower the resolution. The resolution is a measure of the ability of the forecast system to represent the scope of the event. If the line follows the climatological frequency then the forecast system is not able to distinguish between events and non-events. A good forecast system should have good resolution and reliability.

The Brier score can be related to the reliability diagram by partitioning into resolution, reliability and uncertainty. If the probabilities are partitioned into bins as described above, then the Brier score is given by

$$BS = \text{reliability} - \text{resolution} + \text{uncertainty}$$

$$\text{reliability} = \frac{1}{N} F_k \sum_{k=1}^{N_k} \left(\frac{O_k}{F_k} - p_k \right)^2$$

$$\text{resolution} = \frac{1}{N} F_k \sum_{k=1}^{N_k} \left(\frac{O_k}{F_k} - f \right)^2$$

$$\text{uncertainty} = f(1-f)$$

p_k is the forecast probability of the kth category
O_k is the number of occurrences of an event in probability category k
F_k is the number of forecasts in probability category k
N_k is the number of probability categories
N is the number of forecasts
f is the observed climatological frequency of the event

For a good forecast system, the reliability should be close to zero and the resolution large. The uncertainty is dependent on the observed frequency. When the evaluation of precipitation forecasts against radar is performed, there will be a separate curve on the reliability diagram for each square size that is used to generate the probabilities and for each precipitation accumulation threshold. It is expected that the resolution will decrease as the size of the squares becomes larger. The aim is to find the smallest square that gives the best reliability and resolution for each accumulation threshold.

3.3.5 The ROC curve

The Relative Operating Characteristic (ROC) curve is used to examine the relationship between the hit rate and false alarm rate. A good forecast system should produce a large number of hits and a small number of false alarms. Hit rates and false alarm rates are found using a contingency table.

	Observed Yes	Observed No
Forecast Yes	A	B
Forecast No	C	D

Hit rate = A / (A + C)

False alarm rate = B / (B + D)

Ideally, the hit rate should be close to 1 and the false alarm rate close to 0.

A probability forecast for a particular precipitation accumulation threshold can be converted into a contingency table by firstly partitioning the probabilities into thresholds. The hit rate, for each probability threshold, is given by the number of forecast and observed events at and above that threshold divided by the total number of observed events. Similarly, the false alarm rate is given by the number of forecast and non-observed events at and above that probability threshold divided by the total number of non-observed events. The use of the ROC curve is shown in verification by Hand 2002.

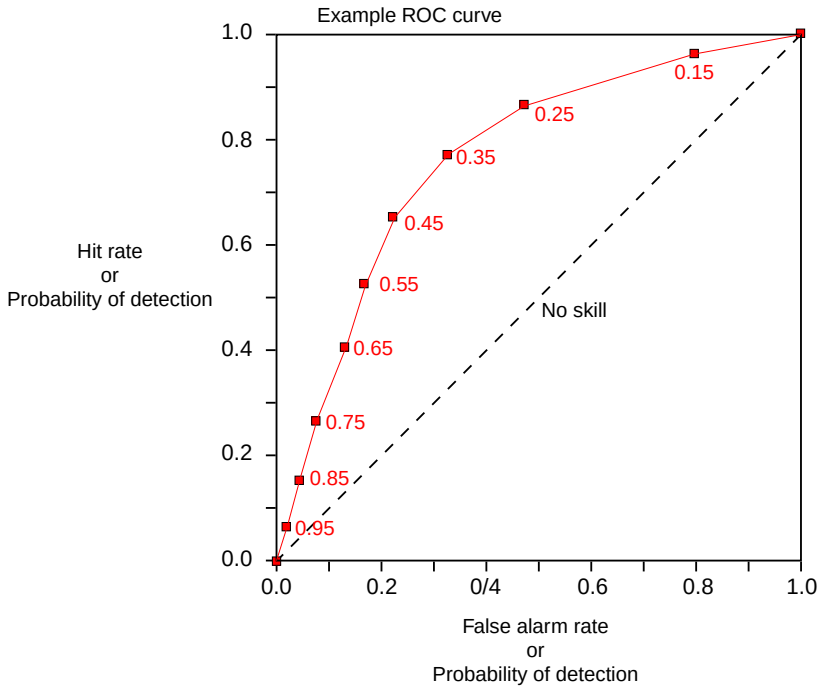

Figure 8. Example of a ROC curve. See text for more information.

Figure 8 shows an example of a ROC curve. The curve starts at (0.0,0.0) because, by definition, the hit rate and false alarm rate will both be 0.0 – a threshold probability of 1 cannot be exceeded, so there can be no hits or false alarms. It ends at (1.0,1.0) because, by definition the hit rate and false alarm rate will both be 1.0 – a threshold probability of 0 includes all occasions, so that all events must be hits and all non-events must be false alarms. A perfectly skilful forecast system will produce a curve that starts at (0.0,0.0) goes to (0.0,1.0) (top left) and finish at (1.0,1.0). A forecast system with no predictive skill would produce a line close to the diagonal. The red line shown lies somewhere in between. A common way to use the ROC curve is to measure the area under the curve: the closer the area to 1.0 the higher the skill. An area of 0.5 indicates no skill. An area > 0.7 is often regarded as the indication of a skilful forecast.

As for the reliability diagram, the ROC curve verification of forecast precipitation accumulations against radar will produce separate curves for each accumulation threshold and for each square size that was used to generate the probabilities.

3.4 How to deal with radar error?

This is included to highlight that it is important rather than to give a satisfactory answer. Radar data have errors associated with them that can be large. Most of the discussion up to this point has ignored that fact. It is going to be difficult to interpret our verification scores if we don't know how truthful 'truth' is and how the uncertainty in the accuracy of the radar data might undermine those scores. This is especially true when using thresholding in the verification. A small error in the magnitude of the radar accumulations might make a huge difference to the number of gridpoints that exceed a particular threshold.

One thing that can be done is to determine what the verification scores will be if the radar fields are compared against themselves with an added error. For example, the radar field can be compared with itself multiplied by a factor of say 2 to take into account the error in magnitude that might exist. The score that is obtained can be interpreted as a 'perfect'

score for the model verification, even though it is not numerically the best score that could be computed. In other words, we can not meaningfully expect the model to be more skilful than the error in the radar data it is measured against. As well as errors in the magnitude of the radar data, there are also spatial errors associated with wind drift and other factors. Again a 'perfect' score could be computed by comparing two radar fields of precipitation rate that are some time interval apart (e.g. 15 minutes).

An alternative approach might be to perform the model verification against radar and also against radar fields that have been modified by an amount equal to an estimated error (positive and negative) in magnitude or location. The difference in scores will give an indication of the sensitivity of the verification to the accuracy of the radar.

The use of frequency thresholds will remove uncertainty due to amplitude error in the score itself and pass this on to the bias. What is meant by a using a frequency threshold is to compare points exceeding say the 90^{th} percentile value rather than a particular accumulation threshold. I.e the top 10% of model accumulation grid squares are compared with the equivalent from radar. It will however, not remove the problem of the spatial error.

It is possible to compute a theoretical score based on the forecast and radar frequencies for a particular event (exceedance of a threshold) for very large spatial scales for some of the verification measures (described later). Given that this can be done, it is also therefore possible to determine the error in the score associated with a given error in the observed frequency. The frequency error can be set to be equivalent to a particular error in the accumulations.

4 Idealised verification examples

This section will show some idealised examples of how the Brier skill score and the odds ratio behave with simple scenarios of observed and forecast precipitation fields. The reliability diagram and ROC curve will not be covered because the sample size (number of gridpoints examined) is not large enough in a single example for them to be meaningful.

All of the idealised examples use a domain of 60 x 60 gridpoints.

The odds ratios were calculated after both the observed and forecast fields were converted into probabilities using the method outlined earlier. The probabilities were converted into fifty contingency tables by using fifty probability thresholds 0.02, 0.04, 0.06 ...0.96, 0.98 to distinguish between a 'yes' or 'no'. The contingency tables were then combined before the odds ratios were computed.
The Brier skill scores were calculated with the forecast fields converted into probabilities and the observed fields remaining as binary 1 for points exceeding the threshold and 0 for points below the threshold.

Both the odds ratios and Brier skill scores are displayed with the score on the y-axis (higher score = more skill) and the square size over which the probabilities were generated along the x-axis. At the origin of the x axis (square length of zero), the probabilities are either 1 or 0 at each point and the scores are the same as would be obtained from a standard point for point verification. Each graph has several curves that show the scores for different displacement distances between the observed and forecast fields.

4.1 Example 1.

The observed and forecast fields are both squares of 20x20 gridpoints representing a region of observed precipitation accumulation above a threshold (see Figure 9)
The squares are initially collocated in the centre of the domain and are then separated horizontally by increments of 4 gridpoints. The separations between the left edge of the 'observed' square and the left edge of the 'forecast' square' are 0 gridlengths (complete overlap), 4, 8, 12, 16 gridlengths (partial overlap) and 20, 24 gridlengths (no overlap).

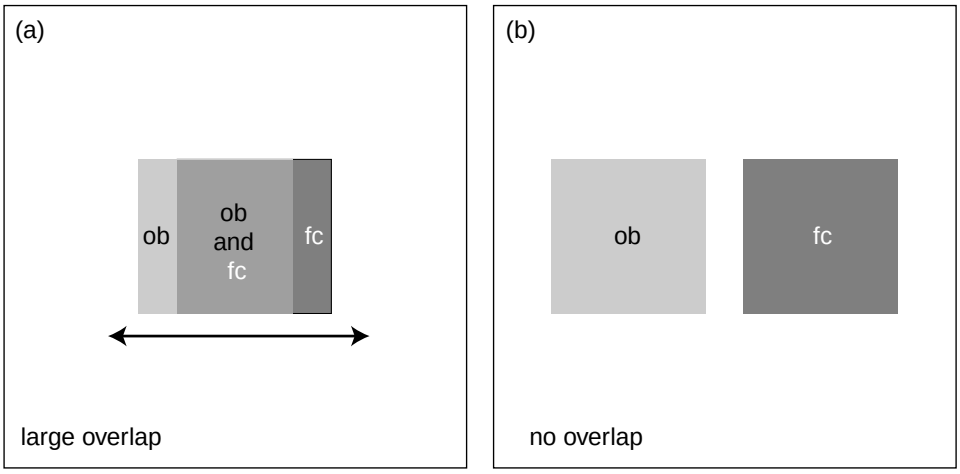

Figure 9. Picture of idealised example 1. (a) A region of precipitation is forecast (light grey square) to be in almost the same location as that observed (darker grey square). (b) The forecast region of precipitation is not in the same location as that observed. Scores are computed for differing forecast displacements from that observed. See text for more information.

The odds ratio graph (Figure 10) shows the following:

1. The scores improve for all displacement errors as the square size used for generating the probabilities is increased. This indicates more skill over larger spatial scales however good or bad the forecast. When the probabilities are computed over even larger squares (not shown), the curves level off and then vanish as the odds ratio can not be computed when the probability fields become too smooth.
2. There is no line for the case when the displacement error is zero (0 gridpoints separation). For such a perfect forecast, the odds ratio will have a value of infinity and can not be plotted. This does present a drawback to using the odds ratio, although in reality, it is very unlikely that a forecast will be absolutely perfect on many occasions. If the score is taken from a large number of events, the problem will be removed.
3. The larger the displacement error in the forecast, the lower the forecast skill for a given sampling area (probability square size). This is what we would hope to see for smaller spatial scales, but when the spatial scales become larger we might hope to see the difference in scores for each curve become less. The two scenarios in which the displacement error is so large there is no overlap between observed and forecast precipitation show no skill over small spatial scales. When there is no overlap, the odds ratio can not be computed unless the length of the square used to generate the probabilities is large enough to bridge the gap. Again, this should not be an issue if the score is computed from a contingency table made up of a large number of events.
4. If it is deemed that an odds ratio score of say 1.0 is sufficiently skilful then the square size over which the probabilities should be computed to give a skilful forecast can be found. These are read from the graph as, square size = 0 gridpoints for 0, 4 and 8 gridpoint separation, square size = 18x18 gridpoints for 12 gridpoints separation,

square size = 32x32 gridpoints for 16 gridpoints separation, square size = 38x38 gridpoints for 20 gridpoints separation.

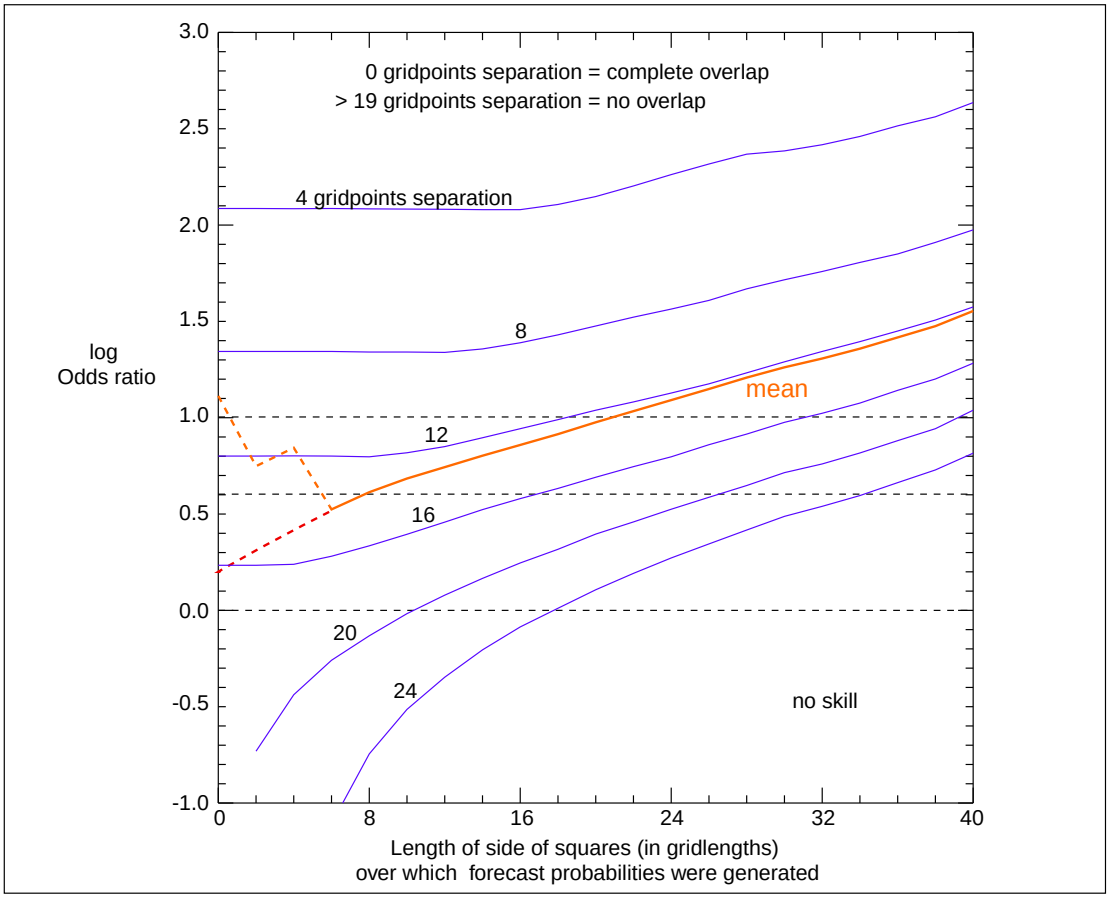

Figure 10. Graph of the log odds ratios against the scale over which probabilities were generated for example 1. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in Figure 9). Scores below 0.0 have no show skill.

The graph of the Brier skill score (Figure 11) shows the following:

1. As with the odds ratios, the larger the displacement error in the forecast (the separation of the forecast and observed precipitation areas), the lower the forecast skill for a given spatial scale (size of square used to generate probabilities).
2. The curves all tend towards the same value as the spatial scale increases. This value can be determined from the frequency of the observed and forecast events. See Appendix A. If the forecast and observed frequencies are the same – as they are in this example (both have 400 points where the event occurs) – then the Brier skill score will tend to a value of 0.5 as the spatial scale increases. The dashed line shows the value that the Brier score will tend towards if the forecast has a frequency of 0 – i.e. the forecast predicts no events occurring.
3. The curves peak at a spatial scale that is dependent on the displacement error in the forecast. The peak in skill occurs at longer spatial scales and becomes flatter as the displacement between forecast and observed becomes larger. When the forecast is perfect (0 gridpoint displacement) the Brier skill score peaks at the gridscale with the perfect score of 1.0; when the displacement is 4 gridsquares the score peaks for forecast probabilities generated over ~9x9 gridpoints; when the displacement is 8 gridsquares the score peaks for forecast probabilities generated over ~18x18 gridpoints – and the trend continues. The peak occurs because, as the spatial scale increases, there is a trade

off between a better fit between the forecast and observed and the tendency of the curves approach a value that depends only on the frequency of the observed and forecast events rather than any spatial error.

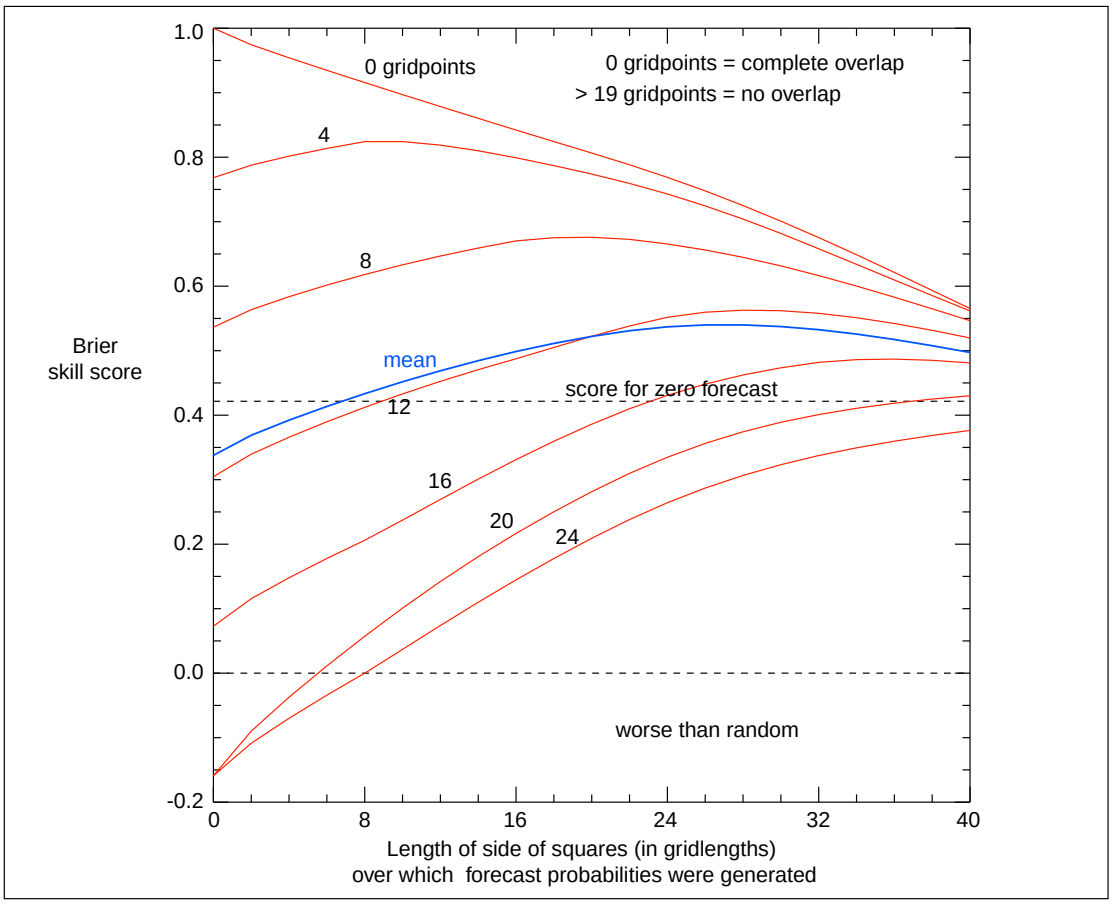

Figure 11. Graph of the Brier skill scores against the scale over which probabilities were generated for example 1. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in Figure 9). Scores below 0.0 show the forecast is worse than a random forecast. The score for a zero forecast is what would be obtained if the forecast produced no precipitation in this example.

4. When there is no overlap between the forecast and observed events (>19 gridpoints separation) the skill for small spatial scales is worse than would be obtained from a random forecast. At the gridscale the forecasts have identically poor scores, but the line with the smaller separation (20 gridpoints) shows better than random skill more quickly as the spatial scale increases.

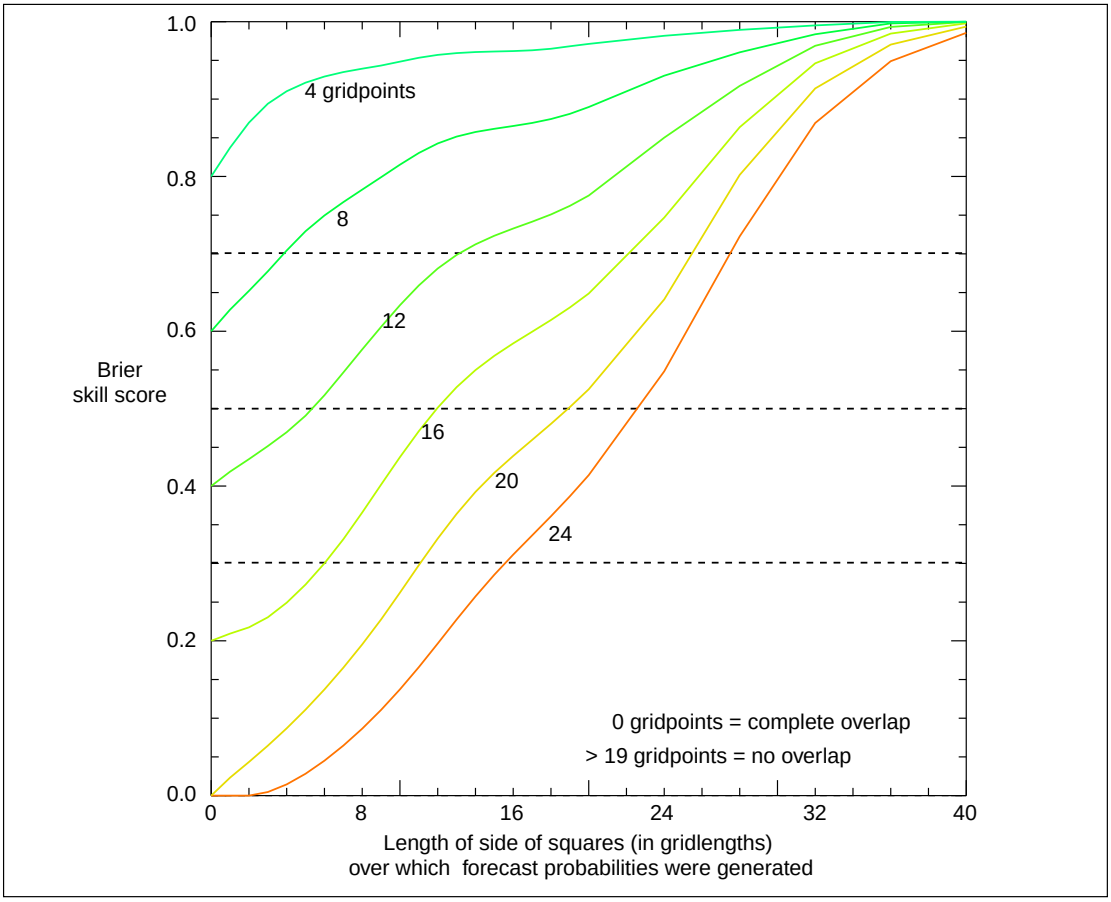

Figure 12. Graph of the fractions skill scores for fractions over square areas against the square sizes over which fractions were generated for example 1. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in Figure 9). Scores of 0.0 show no skill.

The graph of the fractions skill score for comparing fractions over areas shows the following characteristics we would expect.

1. Skill increases with the size of the squares used to compute the fractions.
2. The two scenarios in which the forecast and observed areas have no overlap have a zero score at the gridscale, and the scenario with the greatest spatial error (24 gridpoints) also has no skill for fractions generated over squares of length 2 gridlengths.
3. Since the observed and forecast areas are the same size, the skill score tends to 1.0 for fractions generated over the largest squares.

A curious effect in Figure 12 is the way the lines wiggle. The reason for this is the small size of the domain used and the effect of the edge on this score. The effect is accentuated by these idealised scenarios.

4.2 Example 2

Example 2 contains 2 experiments

Experiment 1: the observed field has a square of 20x20 gridpoints representing a region of observed precipitation accumulation above a threshold and the forecast field has a square of 10x10 gridpoints representing a region of observed precipitation accumulation above the same threshold. See Figure 13(a).
Experiment 2: the observed field has a square of 10x10 gridpoints representing a region of observed precipitation accumulation above a threshold and the forecast field has a square of 20x20 gridpoints representing a region of observed precipitation accumulation above the same threshold. See Figure 13(b).

For both experiments, the squares are initially collocated in the centre of the domain and are then separated horizontally by increments of 4 gridpoints. The separations between the centre of the 'observed' square and the centre of the 'forecast' square' are 0 gridlengths (complete overlap of the smaller square), 4, 8, 12 gridlengths (partial overlap) and 16, 20 gridlengths (no overlap).

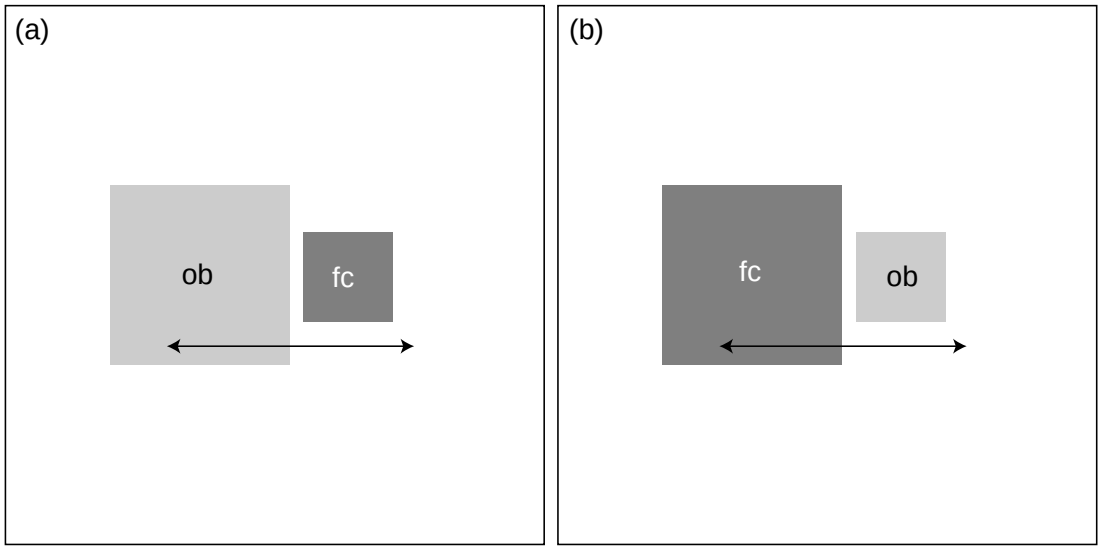

Figure 13. Picture of idealised example 2. (a) Experiment 1. The observed region of precipitation (large light grey square) is 4x the size of the forecast precipitation area. (b) Experiment 2. The forecast region of precipitation (large dark grey square) is 4x the size of the observed precipitation area. In both experiments and the distance between the observed and forecast precipitation areas is varied. See text for more information.

The odds ratio graph is identical for both experiments and has the same general characteristics as seen in example 1. The two experiments produce the same graph because the score is calculated from the probabilities of both the observed and forecast event, which makes the two experiments identical. As in example 1, the curves for 0 gridpoints separation is absent because the score is infinity. In this example, it is also not possible to calculate the odds ratio when the separation is 4 gridlengths.

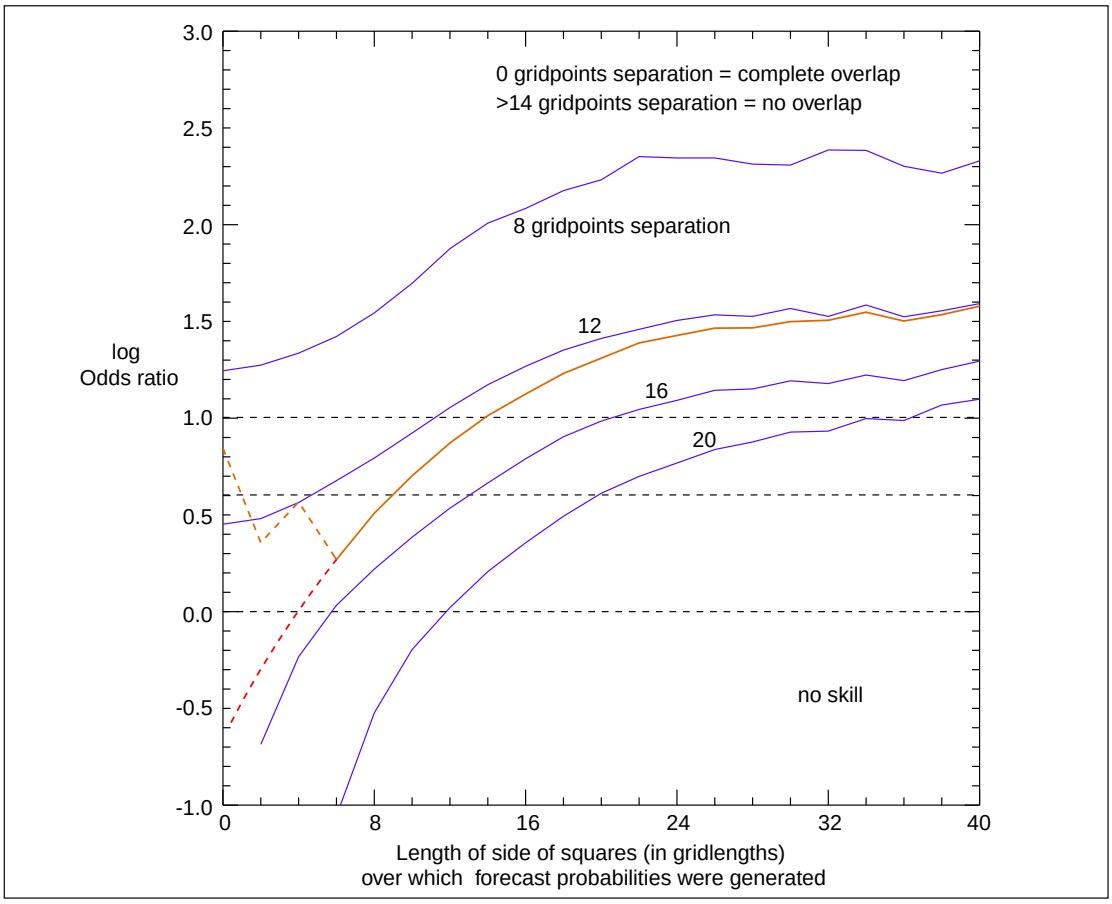

Figure 14. Graph of the log odds ratios against the scale over which probabilities were generated for example 2. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in **Figure 13**). The lines are identical for each experiment (see text). Scores below 0.0 have no show skill.

The Brier skill score curves in Figure 15 are very different between experiment 1 and experiment 2. The set of curves from experiment 1 has higher values and peaks at lower spatial scales than the set of curves from experiment 2. Both sets of curves converge with increasing spatial scale. It is a concern that the curves are different when the spatial differences between the observed and forecast events are the same. The reason why the magnitudes of the curves are different between experiment 1 and experiment 2 is due to the difference in the observed frequency that is used to calculate the random Brier score. When the observed precipitation area is larger the frequency is larger, the random Brier score 2f(1-f) is larger and therefore the Brier skill score is larger. The random forecast with a higher, but still small, frequency will make more incorrect forecasts than the random forecast with a lower frequency. It is therefore easier to improve on the random forecast with the higher frequency. The other problem of the curves peaking at different spatial scales is also due to the difference in frequencies of the observed and forecast events. It is a consequence of comparing forecast probabilities over different spatial scales with binary 1s or 0s for the event occurring or not. This lack of symmetry – i.e. only generating spatial probabilities from one of the fields, means that the comparison in experiment 1 is not the same as that in experiment 2 because of the swap in frequencies. One way to get round this problem is to compare spatial probabilities with spatial probabilities when calculating the Brier skill score. The problem with that approach is that the scores will very quickly improve with spatial scale and reach a value that is dependent on the difference in frequencies, but the this improvement is largely associated with the probability fields both becoming more uniform (loss of resolution), rather than an indication of skill with

increasing spatial scale. A better approach is to make the forecast and observed frequencies the same before doing the verification.

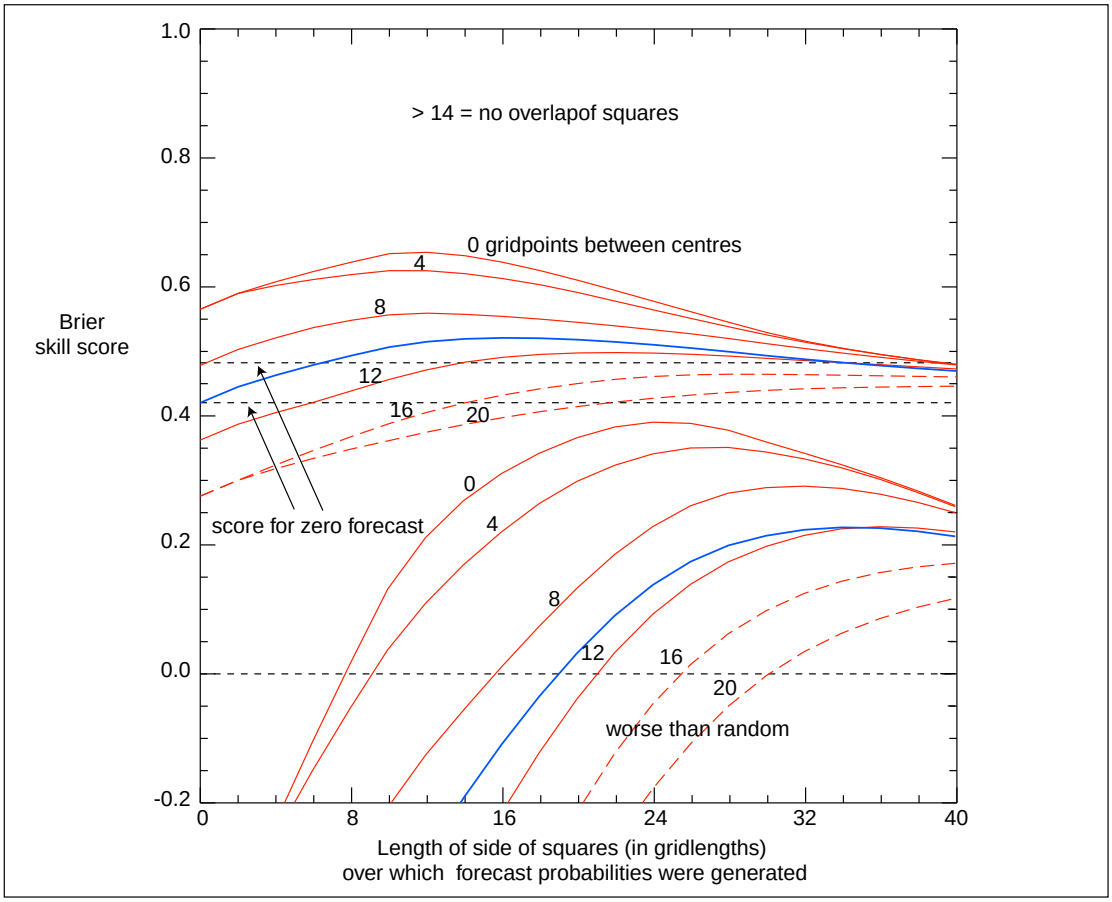

Figure 15. Graph of the Brier skill scores against the scale over which probabilities were generated for example 2. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in Figure 13). The cluster of lines with the higher values is from experiment 1; the others are from experiment 2 (see text). Scores below 0.0 have no show skill.

4.3 Example 3

The observed and forecast fields both have strips of 10x50 gridpoints to represent a region of precipitation accumulation above a threshold associated with a frontal zone (see Figure 16). The strips are initially collocated in the centre of the domain and are then separated horizontally by increments of 2 gridpoints. The separations between the left edge of the 'observed' strip and the left edge of the 'forecast' strip are 0 gridlengths (complete overlap), 2, 4 6, 8 gridlengths (partial overlap) and 10, 12 gridlengths (no overlap).

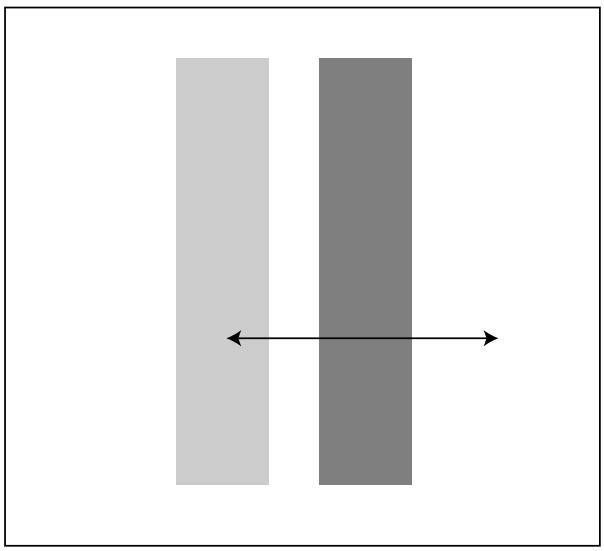

Figure 16. Picture of idealised example 3. The displacement between two identical strips is varied. The strips represent forecast and observed rainfall accumulations from a frontal zone. See text for more information.

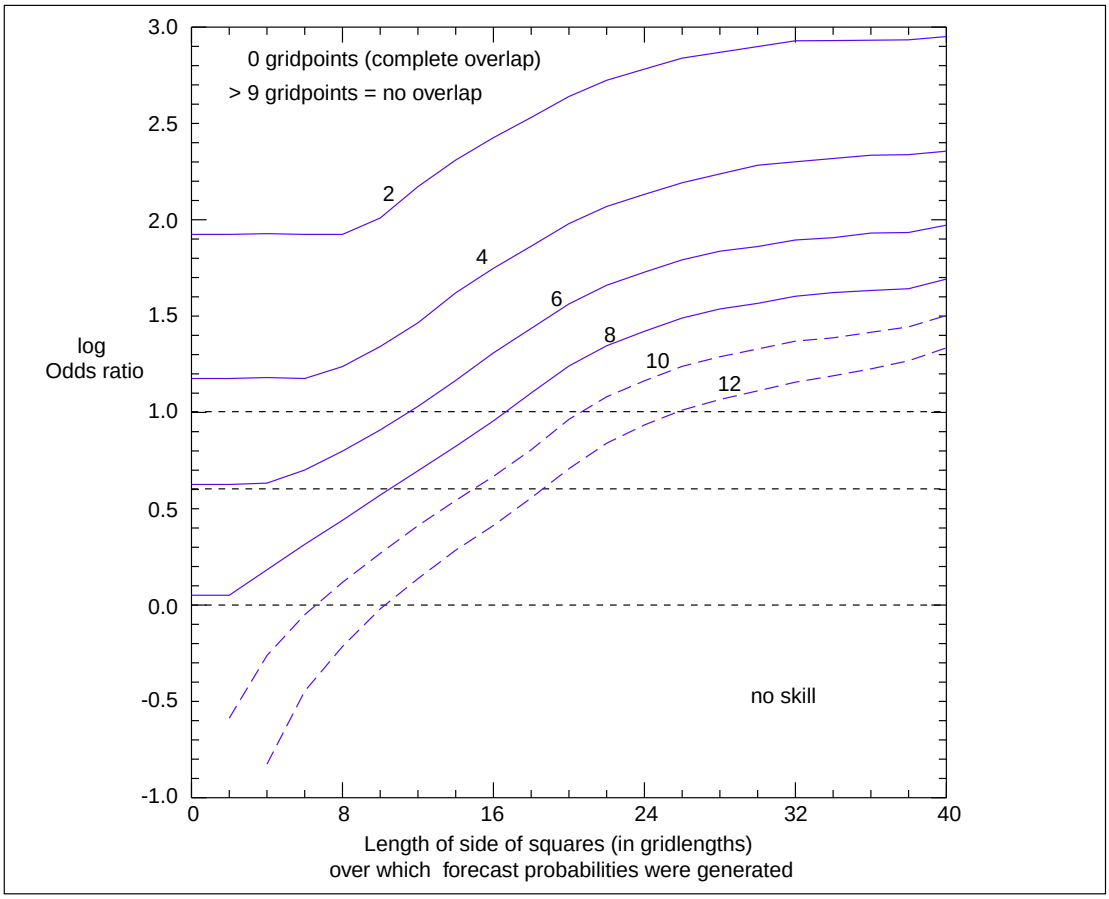

Figure 17. Graph of the log odds ratios against the scale over which probabilities were generated for example 3. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (strips in **Figure 16**). Scores below 0.0 have no show skill.

Figure 17 and Figure 18 show that the odds ratio and Brier skill score curves exhibit the same characteristics as those seen in example 1. They should do, because both examples examine the scores associated with forecast displacement errors in a single region of precipitation that is the same size as the one observed. The differences between example 1 and this one are due to the width of the precipitation area. The square in example 1 is 20 points wide, whereas the strip here is 10 points wide. The narrower strip in this example means that the odds ratio curves increase more sharply with spatial scale and then level off. The Brier skill score curves peak more sharply at smaller spatial scales then converge to the same value more quickly. This behaviour is intuitively reasonable. For a given displacement, the smaller or narrower features require a smaller spatial area to enclose them and give a good score. A further increase in the spatial scale will not result in much or any improvement.

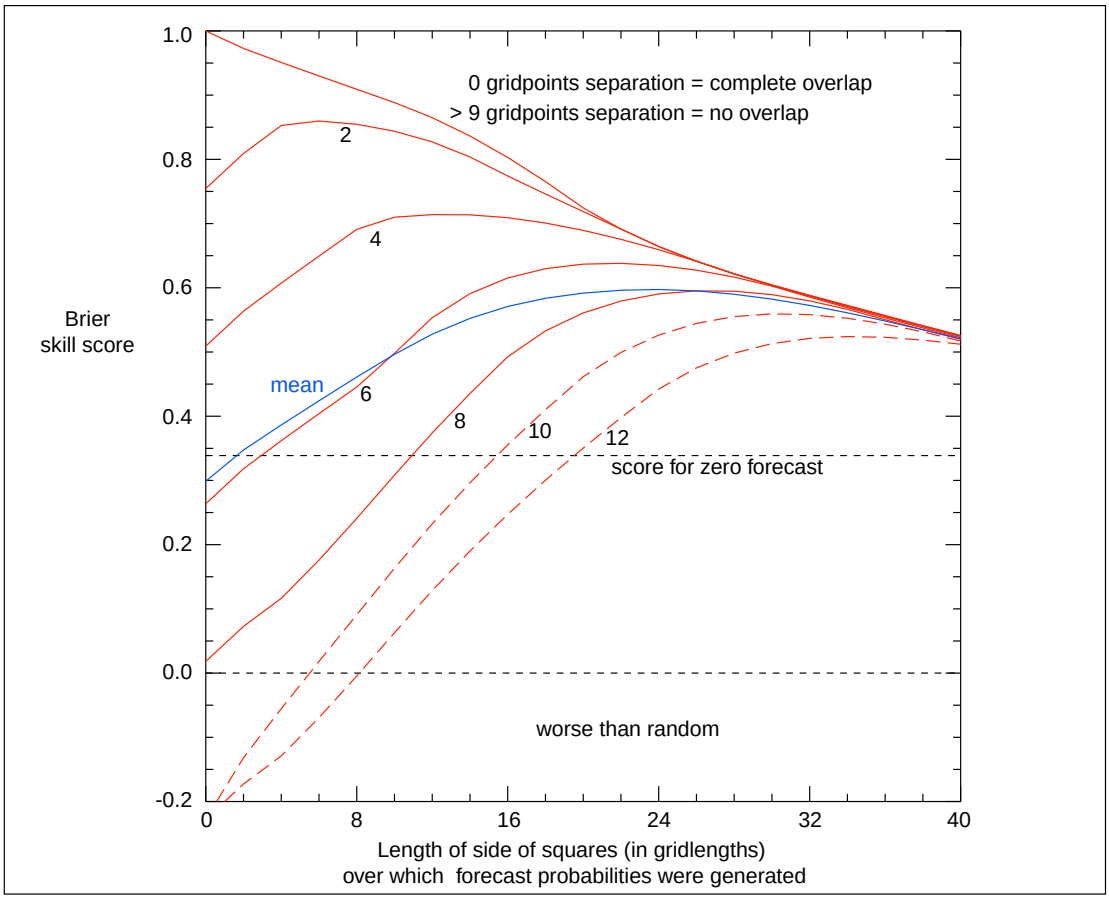

Figure 18. Graph of the Brier skill scores against the scale over which probabilities were generated for example 3. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (strips in **Figure 16**). Scores below 0.0 show the forecast is worse than a random forecast. The score for a zero forecast is what would be obtained if the forecast produced no precipitation in this example.

4.4 Example 4

The observed and forecast fields both have squares of 5x5 gridpoints representing precipitation, 5x5 gridpoints apart, tiled across the domain (see Figure 19). This pattern is meant to represent 'scattered' precipitation areas. The squares are initially collocated so that forecast 'showers' are in the same positions as observed 'showers' and then the

forecast pattern is separated diagonally from the observed by increments of 1 gridpoint in x and y. At 5 gridpoints separation there is no overlap – no forecast showers are in the same place as observed showers.

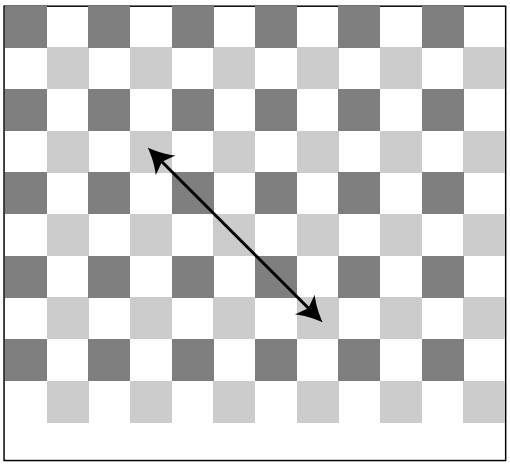

Figure 19. Picture of idealised example 4. (a) Dark/light grey squares represent observed/forecast precipitation areas. The forecast squares are moved to give situations with differing positional agreement between the observed and forecast to provide the lines in **Figure 20** and Figure 21. See text for more information.

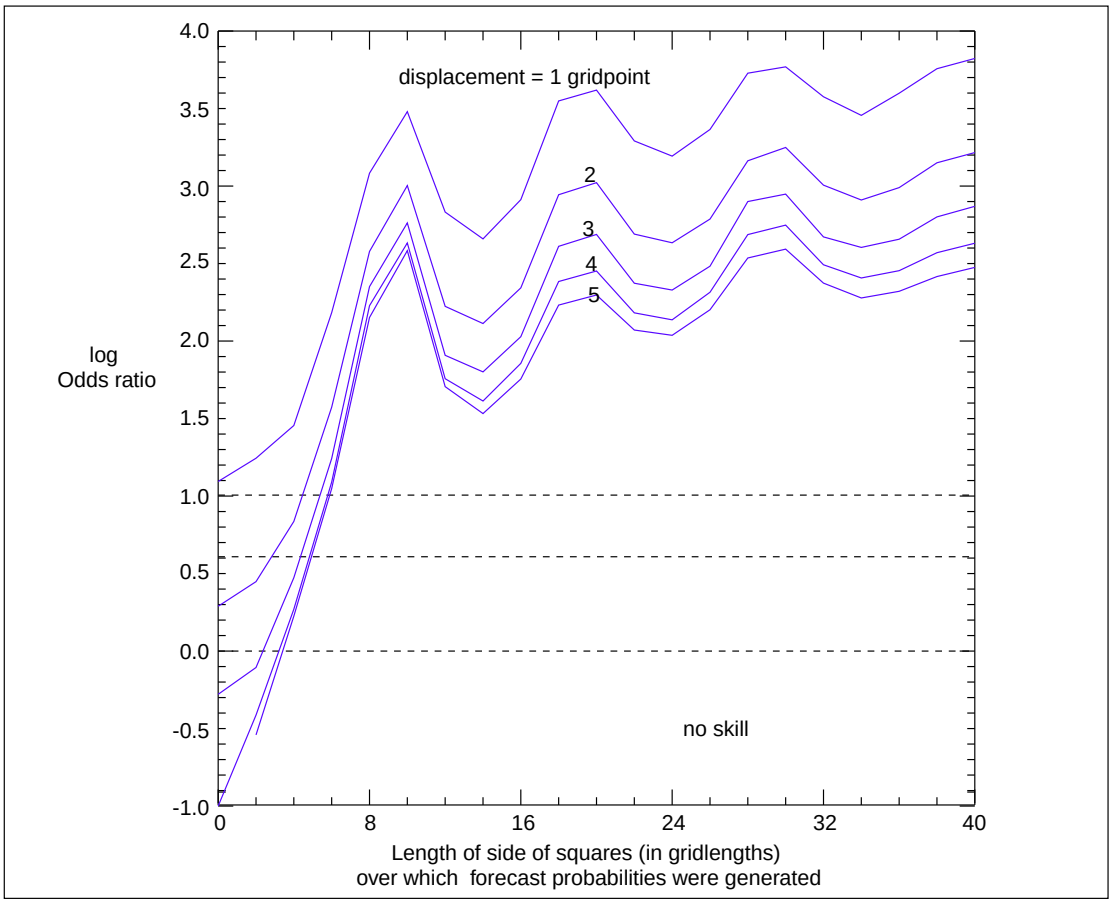

Figure 20. Graph of the log odds ratios against the scale over which probabilities were generated for example 4. Each line shows the scores for a different positional agreement between the forecast and observed precipitation areas (squares in **Figure 19**). Scores below 0.0 have no show skill.

The odds ratio curves show some similarities and differences to the other examples. As before, there is no curve for the perfect forecast. All the curves show a rapid increase in odds ratio with spatial scale up to 10x10km squares and then there is a more gradual oscillating increase. Given the setup of this example it is encouraging that there is a peak at 10km. It is also encouraging that the curves have the rapid increase in score at roughly the same spatial scale. The concern is that even at the larger spatial scales when all the forecasts should be comparable, the forecasts that are better at the smaller scales are still better. Nevertheless, it does seem to show that when the forecast has the correct pattern, but is not necessarily positioning individual elements in the correct place, there are larger spatial scales over which the forecast can be considered skilful by this measure.

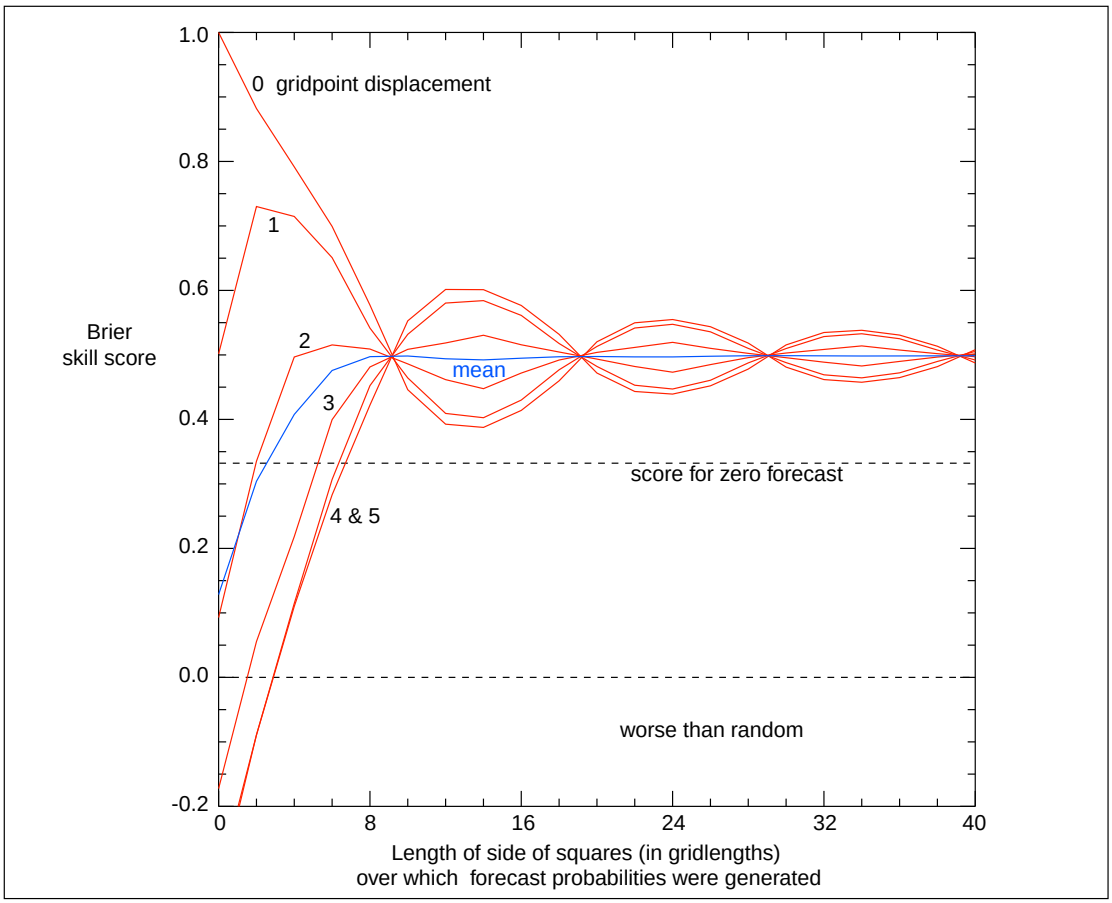

Figure 21. Graph of the Brier skill scores against the scale over which probabilities were generated for example 4. Each line shows the scores for a different positional difference between the forecast and observed precipitation areas (squares in **Figure 19**). Scores below 0.0 show the forecast is worse than a random forecast. The score for a zero forecast is what would be obtained if the forecast produced no precipitation in this example.

The Brier skill score curves have interesting characteristics. All of the curves converge in an oscillatory manner to 0.5 (the theoretical value) at larger spatial scales. At small spatial scales the forecasts which are closest to the observed have much better scores. The perfect forecast has a perfect score of 1.0 at the gridlength scale (0 length square). The forecast with a 1-gridpoint displacement has a peak at a spatial scale of 2 gridlengths (probabilities

computed over 3x3 points). The forecasts with a 4 or 5 gridpoint displacement have less skill than a random forecast at very small spatial scales but are comparable to the others when the spatial scale exceeds 8 to 10 gridlengths. It is a useful characteristic of the score that all the curves tend to approach the same value with a reduced amplitude oscillation. It shows that the different displacement errors give very different skill scores for small spatial scales but are essentially the same over larger scales. This is intuitively what should be expected – and what we would hope to see, for this kind of scenario. The oscillation is a concern though – it can make the best forecast appear the worst at some spatial scales. Another feature of the Brier skill score in this case is that the value of 0.5 the curves converge towards is significantly higher than the value that would be obtained if the forecast had predicted no events. In examples 1 and 2 this is not the case. The reason is simply that the observed frequency is higher in this example and example 3 (the front) – i.e. there are more precipitation points than in examples 1 and 2, and any forecast that does not attempt to predict anything will do worse in this situation than in the two with fewer precipitation points. It confirms what we already know, that the rarer the event the better a forecast system that predicts nothing will be.

4.5 Example 5

In this example there is only one scenario – the one shown in Figure 22. The observed and forecast fields both have squares of 10x10 gridpoints representing regions of precipitation accumulation above a threshold. The left-most squares completely overlap, the squares in the middle of the domain are adjacent without overlap and the right-most squares are separated by 10 gridlengths.

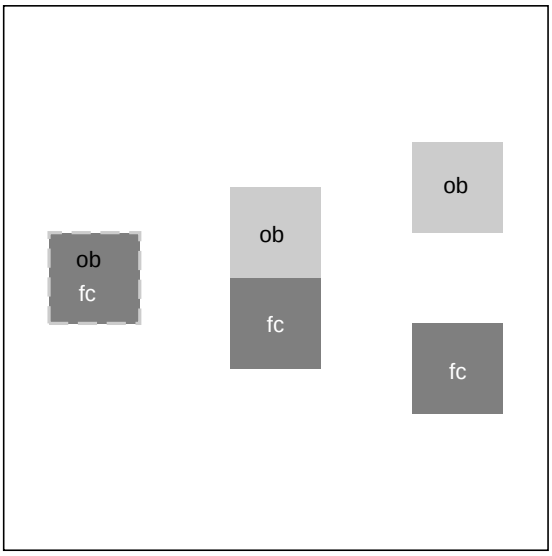

Figure 22. Picture of idealised example 5. See text for more information.

The odds ratio curve in Figure 23 shows a continuous improvement with increasing spatial scale and has a reasonable score of ~0.8 at the gridscale. If 1.0 is regarded as a threshold score for a sufficiently skilful forecast then the spatial scale over which the forecast is that skilful is around 8 gridlengths. As a comparison, a score of ~1.0 would be obtained at the gridscale (x=0) if there was an overlap of 3 gridrows (30 points) between the middle forecast and observed squares and nothing else was changed. The gridscale score for a scenario in which all the square precipitation areas were 20x20 points rather than 10x10, but the pattern was identical in terms of overlap to that shown, is 0.38 (and 1.45 for 5x5

squares). For large spatial scales comparable to the size of the domain, the score continues to increase to values > 3.5 (not shown in graph).

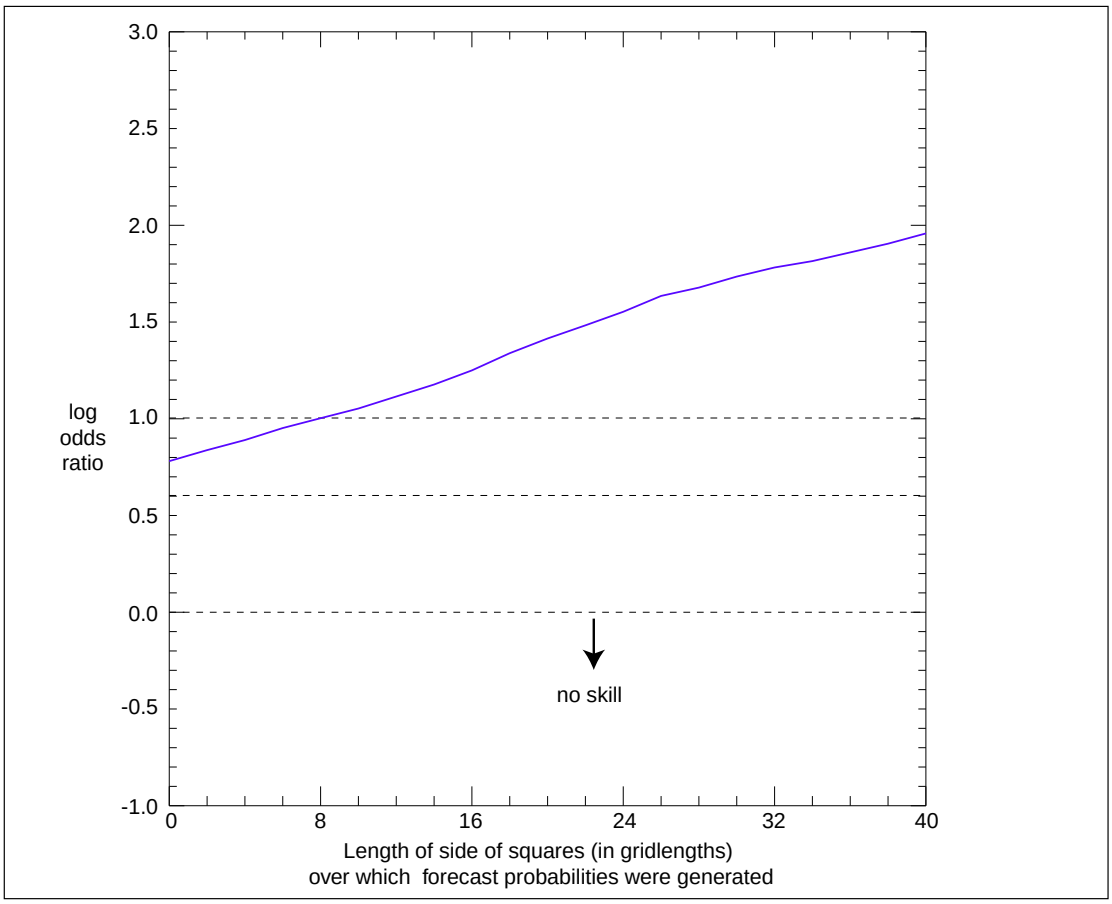

Figure 23. Graph of the log odds ratios against the scale over which probabilities were generated for example 5 (see **Figure 22**). Scores below 0.0 have no show skill.

The Brier skill score curve in Figure 24 shows an increase in skill from the gridscale to a spatial scale of ~12 to 14 gridlengths after which the forecast skill is mostly independent of spatial scale. Beyond 10 gridlengths, the forecast has more skill than a forecast that does not predict anything – and in this case it is difficult to beat a zero forecast because the observed frequency is low. The graph indicates that for this particular forecast, the best choice of square size to compute the probabilities over is around 12 gridlengths. Any larger than that and the resolution of the forecast is being lost as the probabilities get more smoothed out and less useful, for no extra gain in skill. In this example the skill never reaches a very high value over any spatial scale (never greater than the asymptote of 0.5) because the forecast at the gridscale is not particularly good. However, the forecast at all scales is still considerably better than a random forecast would be.

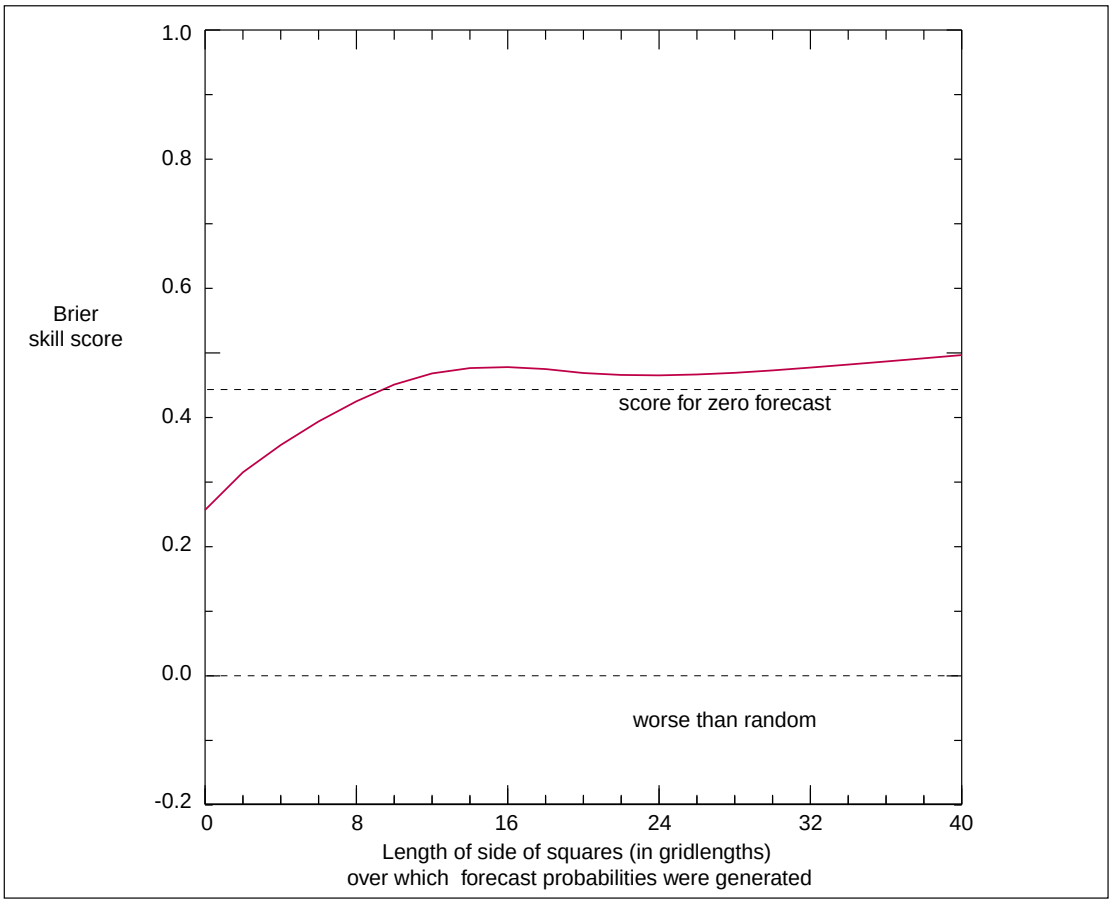

Figure 24. Graph of the Brier skill scores against the scale over which probabilities were generated for example 5. Scores below 0.0 show the forecast is worse than a random forecast. The score for a zero forecast is what would be obtained if the forecast produced no precipitation in this example.

Figure 25 shows how the fraction skill score used for comparing fractions exceeding thresholds behaves for this example when the fractions are computed over different sized squares. The line tends to 1.0 for fractions generated over the larger squares, as expected when the observed and forecast areas are the same. As the square sizes get larger the skill increases rapidly up around 12 gridlengths, then flattens out. It then increases more gradually until around 20 gridlengths, then increases rapidly again. This does not seem unreasonable given the scenario shown and matches the peak and trough seen in the Brier skill score (Figure 24).

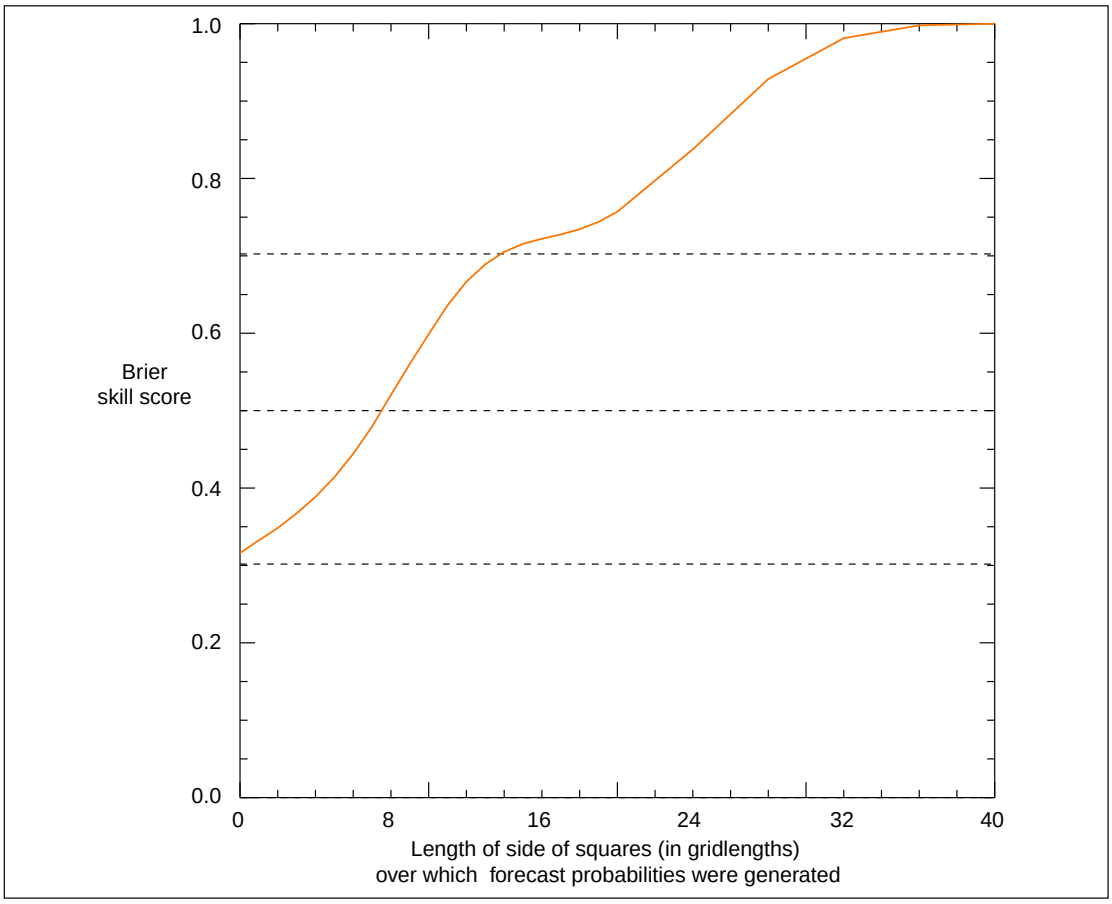

Figure 25. Graph of the fractions skill score for area fractions against the size of square over which fractions were generated for example 5. A score of 0.0 indicates the forecast has no skill (at that scale).

4.5.1 Conclusions from the idealised examples

Used together, the Brier skill scores, the fractions skill score and odds ratios computed from fractions/probabilities generated over different sized sampling areas can give a useful indication of the skill of a forecast system over different spatial scales. The scales over which the scores indicate skill do agree with what would be expected from visual inspection. However, the scores should probably be interpreted in two different ways. For the odds ratios, the actual value is useful and it may be appropriate to define the spatial scale over which a forecast is sufficiently accurate to be when the odds ratio is greater than some threshold (say 1.0). For the Brier skill score, the spatial scale at which a forecast is sufficiently accurate is found from the peak in the curve or where the curve tends to level off – the actual value is not so useful. The comparison between the Brier skill scores obtained and the theoretical scores for a very large spatial scale, a zero forecast and a random forecast provides additional information about forecast performance.

The absolute value of the skill score used for comparing fractions over areas is meaningful, especially if frequency thresholds are used. However, even when accumulation thresholds are used, the behaviour of the score is in good agreement with what the human eye qualitatively expects. It also has the advantage of giving a score of zero for a forecast that does not try (i.e. produces no rain)

The odds ratio can have a problem with zero values in the contingency table. This will crop up when either the forecast is so good it never has a false alarm (unlikely) or the

forecast is so poor it never gets a correct prediction (possible for rare events). Either way, the problem is alleviated if the contingency table is constructed over a large number of forecasts – and statistics over a large number of forecasts is what we want anyway. The question of how to convert the probabilities into categorical scores remains a problem, but these examples do seem to show that the method of using a large number of probability thresholds does produce sensible and useful results.

A useful interpretation of the Brier skill score can be adversely affected by a difference in frequency between the forecast and observed events. The solution is to generate forecast probabilities from frequency thresholds rather than accumulation thresholds. The drawback is that the accumulations will then not match, but upside is that the mismatch is a measure of the bias, which is required in any case to get meaningful verification.

5 Intended assessment

A set of high-resolution analyses and forecasts of precipitation accumulations will be verified against radar accumulations. The bulk of the model simulations that will be evaluated will come from experiments performed within the High ResolutionTrial Model (HRTM) project (Lean 2003). These simulations have been run with gridlengths of 12, 4 and 1km. The maximum length of the forecasts will be 7 hours from the nominal T+0 'analysis' time. The baseline runs use 3DVAR and MOPS data assimilation at 12 and 4km and no data assimilation as such in the 1km model, but increments from the 4km 3DVAR are added.

The model and radar data will be interpolated or averaged to the same 5km-gridlength domain. The domain will be the same size as used for the 1km-gridlength runs and have 60x60 gridpopints. Hourly and six hourly precipitation accumulations will be compared. Area fractional exceedances (can be called probabilities) will be computed for precipitation accumulations on the 5km-gridlength-verification domain using squares up to 400x400km. Area fractions from model forecasts will be compared with area fractions from radar and with point 1 or 0 probabilities from radar. The accumulation thresholds are likely to come from the following 0.01, 0.2, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 24.0, 32.0 and 48.0mm. Percentile thresholds are likely come from the following 50, 75, 90, 95, 99 and 99.9%.

The use of a 5km-gridlength domain may have a tendency to show an under-forecasting of the 12km model for the higher accumulation thresholds. For that reason, it may be useful to also perform the verification on the same domain, but with a 15 and 45 km gridlength to make a fairer comparison. However, this will remove the highest accumulations from the verification, so it should only be done in addition to the 5km-gridlength assessment.

The verification scores used to assess the forecast skill over all events on different spatial scales using probabilities will be: -

1. The odds ratio
2. The Brier skill score
3. Fraction skill score
4. The reliability diagram (and reliability and resolution scores)
5. The ROC curve (and area underneath the ROC curve)
6. Discrimination. This is the mean forecast probability for all occasions when an event occurred minus the mean forecast probability for all occasions when an event did not occur. The range is +1 (best) to –1 (worst) and gives a snapshot view of the performance of the model based on the occurrence or non-occurrence of the event.

The other verification scores used will be: -

1. The bias in frequencies at all thresholds
2. The equitable threat score on a gridpoint by gridpoint basis
3. The bias in total accumulations

The calculation of the bias is essential in conjunction with the probability-based scores to give a balanced assessment of forecast skill. Other scores may be introduced.

Once the scores have been computed the following questions can be addressed

1. What is the impact of data assimilation in the 4 and 1km gridlength models on 6-hour accumulation forecasts?
2. How does the skill of the forecasts vary with spatial scale?
3. How does the skill of 1-hour accumulation forecasts vary with spatial scale, forecast length and use of data assimilation?
4. What are the most appropriate scales to generate probabilities over in order to produce useful output diagnostics?
5. How do the different resolution models compare in the aspects listed above?
6. What threshold probability provides the best guidance

The ability to answer these questions should prove very informative, but it must be remembered that the number of forecasts being examined at this stage is small and does not cover anything like the full spectrum of precipitation events. It is a start, but further statistics from a large number of events will be required. The framework to do this has now been established and should form the basis of the verification of precipitation forecasts from future high-resolution trials.

6 Appendix A

Brier score $BS = \frac{1}{N} \sum_{j=1}^{N} (p_j - o_j)^2$

$0 \leq p_j \leq 1$ probabilistic forecast
$o_j = 1$ if event occurs
$o_j = 0$ if event doesn't occur
N = Total number of gridpoints

If the area over which the probabilities are generated becomes very large then
$p_j \longrightarrow f_{fc}$ where f_{fc} = the forecast frequency

$$BS = \frac{1}{N} \sum_{j=1}^{N_1} (f_{fc} - 1)^2 + \frac{1}{N} \sum_{j=1}^{N_0} (f_{fc} - 0)^2$$

N_1 = Number of observed events
N_0 = Number of observed non-events

$$BS = f (f_{fc} - 1)^2 + (1 - f) f_{fc}^2$$

f = The observed frequency = $\frac{N_1}{N}$

The Brier skill score (BSS) is given by

$$BSS = 1 - \frac{BS}{BS_{reference}}$$

Where $BS_{reference}$ is the reference Brier score given by

$$BS_{reference} = 2f (1 - f)$$

It is the Brier score for a random forecast with the observed frequency

$$BSS = 1 - \frac{f (f_{fc} - 1) + (1 - f) f_{fc}^2}{2f (1 - f)}$$

If the forecast and observed fields have the same frequency (no bias)
$f = f_{fc}$

$$BSS = 1 - \frac{f^3 - 2f^2 + f + f^2 - f^3}{2f (1 - f)} = 1 - \frac{f - f^2}{2(f - f^2)} = 0.5$$

7 Appendix B

$$\text{Brier score } BS_{\text{fractions}} = \frac{1}{N}\sum_{j=1}^{N}(p_j - o_j)^2$$

$0 \leqslant p_j \leqslant 1$ forecast area fraction

$0 \leqslant o_j \leqslant 1$ observed area fraction

N = Total number of gridpoints

At the limit for fractions computed over very large areas

$p_j \longrightarrow f_{fc}$ where f_{fc} = the forecast frequency

$o_j \longrightarrow f_{ob}$ where f_{ob} = the observed frequency

therefore

$$BS_{\text{fractions}} = (f_{fc} - f_{ob})^2$$

and

$$\text{Worst case } BS_{\text{fractions}} = f_{fc}^2 + f_{ob}^2$$

(This is the score that would be obtained if there was no overlap between forecast and observed rainfall fractions)

The skill score relative to the worst case score for area fractions computed over very large areas (squares approaching the size of the domain) is given by:

$$\text{Skill score} = 1 - \frac{BS_{\text{fractions}}}{\text{Worst case } BS_{\text{fractions}}}$$

$$\text{Skill score} = 1 - \frac{(f_{fc} - f_{ob})^2}{f_{fc}^2 + f_{ob}^2} = \frac{2 f_{fc} f_{ob}}{f_{fc}^2 + f_{ob}^2}$$

Therefore for fractions computed over squares approaching the size of the domain

If either f_{fc} or f_{ob} = 0.0, the skill score = 0.0

If $f_{fc} = f_{ob}$, the skill score = 1.0

8 References

Golding, B.W. 1998: Nimrod: A system for generating automated very short range forecasts., *Meteorol. Appl.* 5, pp1-16

Hand, W.H. 2002: The Met Office Convection Diagnosis Scheme., *Meteorol. Appl.* 9, pp69-83

Jones, C.D. and Macpherson, B. 1997: A Latent Heat Nudging scheme for the Assimilation of Precipitation Data into an operational Mesoscale Model. *Meteorol. Appl.*, 4, 269-277.

Lean, H. W., 2003: High Resolution Trial Model project, results following the completion of stage 3. Available on request.

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne and F. W. Saunders. 2000: The Met. Office Global 3-Dimensional Variational Data Assimilation Scheme. *Quart. J. Roy. Met. Soc.*, 126, 2991-3012.

Macpherson, B. 1996: The impact of MOPS moisture data in the UK Meteorological Office Mesoscale Data Assimilation Scheme., *Mon Wea Rev*, 124., 1746-1766

Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, 12, 595-600.

Roberts, N. M., 2003b: The impact of a change to the use of the convection scheme in high-resolution simulations of convective events. *JCMM internal report.,* 142. *(NWP Technical Report.,* 407*)*

Roberts, N. M., 2003c: Precipitation diagnostics for a high resolution modelling system. *JCMM internal report.,* 145. *(NWP Technical Report.,* 423*)*

Smith, R.N.B., 1990. A scheme for predicting layer clouds and their water content in a general circulation model. *Q.J.R. Meterol. Soc.*, 116, 435-460

Stephenson, D.B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, 15, 221-232.